A Survey on MR-MNBC: MAX-REL based Feature Selection for the Multi-Relational Bayesian Belief Network

Disha Sheth[1] Premal Patel[2]
[1]Student of M.E [2]Head of Department & PH.D
[1,2]Department of Computer Engineering
[1,2]Ipcowala Institute of Engineering and Technology

Abstract— High dimensional data often contains irrelevant features that reduce the accuracy of data mining techniques and slow down the process and it is hard to interpret so the Feature selection has become an active area in the data mining. Feature selection selects the relevant subset of attributes, provides the better accuracy and improves the comprehensibility of the models. Feature selection also defying the curse of dimensionality to improve prediction performance. We will propose the method which is based on Feature selection as a preprocessing task of MRDM on probabilistic model. We analyzed our algorithm over large pkdd dataset and get the better accuracy compare to the existing methods.

Key words: Multi-relational classification, Tuple ID Propagation, Semantic Relationship Graph, Bayesian Belief Network, Feature Selection

I. INTRODUCTION

Data mining is the technique to extraction of valuable knowledge from large amounts of data. Data mining is the process of discovering knowledge from data [1]. Data mining has a variety of fields which provides the different tools and the techniques for handling the large database. To deal with this problem, one either constructs a single table by Propositionalization, or uses a Multi-Relational Data Mining algorithm. Conventionally, many classification approaches can only be applied to a single relation. When performing these approaches on multirelational data, it often requires transferring data into a single table by flattening and feature construction, which is known as Propositionalization. MRDM approaches have been successfully applied in the area of bioinformatics. MRDM allowing applying directly in the data mining in multiple tables [2]. To avoid the expensive joining operations and semantic losses we used the MRDM technique.

This paper focuses some of the application areas of MRDM. The task of classification is concerned with predicting the value of one field from the values of other field. In recent years, there has been growing interest in multi-relational classification research and application, which address the difficulties in dealing with large relation search space, complex relationships between relations, and a daunting number of attributes involved. Most structured data is stored in relational databases, which is stored in multiple relations by their characters. However, many of these methods are heuristic, so flatten may cause some problems such as time consuming and statistical skew on data. Multi-relational data mining (MRDM) has been successfully applied in a variety of areas, such as marketing, sales, finance, fraud detection, and natural sciences [1]. Multi-Relational data mining [3] looks for patterns that involve multiple relations in a relational database; its main difference with traditional data mining approaches is that it does not need to transform the data into a single table. This paper discusses a Bayesian Belief Network based approach for multi relational data mining that builds a probabilistic model directly from multiple tables and also exploits relation between tables as well.

II. DATA PREPROCESSING

Data may be in different form as it comes from different sources, it may consist of noisy data, irrelevant attributes, missing data etc. Data needs to be preprocessed before applying any kind of data mining algorithm which is done using following steps:

A. Data Integration:

If the data to be mined comes from different sources data needs to be integrated which involves removing inconsistencies in names of attributes or attribute value names between data sets of different sources.

B. Data Cleaning:

This step may involve detecting and correcting errors in the data, filling in missing values, etc.

C. Discretization:

When the data mining algorithm cannot cope with continuous attributes, discretization needs to be applied.

D. Attribute Selection:

All attributes are not relevant so for selecting a subset of attributes relevant for mining, among all original attributes, attribute selection is required.

III. INDUCTIVE LOGIC PROGRAMMING (ILP)

For multi-relational classification, Inductive logic programming has attracted many researchers in early days of MRDM [1], [2]. Multi-relational classifier becomes more precise if we find relevant features in non-target relation that differentiate target tuples. Target and non-target relations are connected via multiple join paths. Complex schema needs to search a large number of join paths. It is very time consuming to identify good features and repeatedly explore and link up the relations along different join paths and need to evaluate it. Although, Inductive logic programming approaches are efficient and provides good classification accuracy, but they are not scalable. Many researchers are figuring out on how to build scalable and efficient ILP algorithms. Building a decision tree from stored data is one of the approaches. Efficiency can be improved by evaluating a bunch of quarries that have common prefaces [3]. The main drawback is that the query should be known. To overcome this problem, CrossMine [4] was developed.

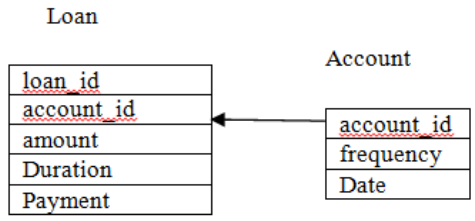

Fig. 1: Loan table and Account table relationship in PKDD cup99 dataset [8]

IV. SEMANTIC RELATIONSHIP GRAPH

For a classification task in a multi-relational database, there is usually one table containing the class label attribute. We call this table as target table, and call the class label attribute as target attribute. Apart from the target table, there are usually many other tables linked to the target table directly or indirectly through arbitrarily long chains of joins. In order to represent this kind of relationship between tables, we use a graph, which is called a semantic relationship graph[8].

A. Definition: (Semantic Relationship Graph):

Semantic Relationship Graph is a directed acyclic graph SRG (*V*, *E*, *W*), where *V* is a set of vertices, each of which corresponding to a table in the database. *E* is a set of directed edges, and an edge (*v*, *w*) means table *w* can be linked to table *v* by directly joining these two tables. W is a set of attributes, each of which links two tables. We call this kind of attribute link attribute. Each edge of the semantic relationship graph represents one of the following two relationships between table's *v* and *w*:

(1) Primary-key to foreign-key relationship, indicating that table w contains foreign-key referring to primary-key in table v.
(2) Foreign-key to primary-key relationship, indicating that table v contains foreign-key referring to primary-key in table w

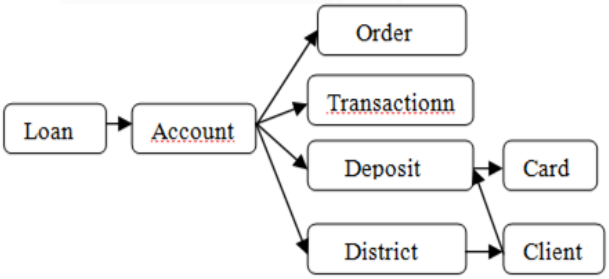

Fig. 2: Semantic Relationship graph for financial data[5]

V. TUPLE ID PROPAGATION

Tuple ID propagation [8] is a method for virtually joining non-target relations with the target relation. It is flexible and efficient method and it avoids the high cost of physical join. Suppose the primary key of the target relation is an attribute of integers, which represent the IDs of the target tuples. We use the ID of each target tuple to represent that tuple. This process takes only small amount of time and space compared to the physical joins used by the existing classifiers and it will boost up the effectiveness of the multi-relational classification techniques. Tuple ID propagation approach reveal to search in the relational database and which is observed that less costly than physical joins in both time and space.

Loan					
loan_ID	account_ID	amount	duration	payment	class
1	124	1000	12	120	+
2	124	4000	12	350	+
3	108	10000	24	500	-
4	45	12000	36	400	-
5	45	2000	24	90	+

Account				
account_ID	frequency	date	IDsets	class labels
124	monthly	960227	1,2	2+, 0-
108	weekly	950923	3	0+, 1-
45	monthly	941209	4,5	1+, 1-
67	weekly	950101	-	0+, 0-

Fig. 3: Example of Tuple ID Propagation [1]

VI. BAYESIAN BELIEF NETWORKS

A Bayesian network (BN)[9] consists of a directed, acyclic graph and a probability distribution for each node in that graph given its immediate predecessors. A Bayes Network Classifier is based on a bayesian network which represents a joint probability distribution over a set of categorical attributes. It consists of two parts, the directed acyclic graph G consisting of nodes and arcs and the conditional probability tables. The nodes represent attributes whereas the arcs indicate direct dependencies. The density of the arcs in a BN is one measure of its complexity. Sparse BNs can represent simple probabilistic models (e.g., naïve Bayes models and hidden Markov models), whereas dense BNs can capture highly complex models. Thus, BNs provide a flexible method for probabilistic modeling [5].

VII. FEATURE SELECTION

In the past thirty years, the dimensionality of the data involved in machine learning and data mining tasks has increased explosively. Data with extremely high dimensionality has presented serious challenges to existing learning methods [9], i.e., the curse of dimensionality. With the presence of a large number of features, a learning model tends to overfit, resulting in their performance degenerates. To address the problem of the curse of dimensionality, dimensionality reduction techniques have been studied, which is an important branch in the machine learning and data mining research area. Feature selection is a widely employed technique for reducing dimensionality among practitioners. It aims to choose a small subset of the relevant features from the original ones according to certain relevance evaluation criterion, which usually leads to better learning performance, lower computational cost, and better model interpretability. According to whether the training set is labeled or not, feature selection algorithms can be categorized into supervised [6], unsupervised [1] and semi-supervised feature selection [1]. Supervised feature selection methods can further be broadly categorized into filter models [8], wrapper models [8] and embedded models[9].

The filter model separates feature selection from classifier learning so that the bias of a learning algorithm does not interact with the bias of a feature selection algorithm. It relies on measures of the general characteristics

of the training data such as distance, consistency, dependency, information, and correlation. Relief [6], Fisher score [1] and Information Gain based methods [2] are among the most representative algorithms of the filter model.

The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the quality of selected features. These methods are prohibitively expensive to run for data with a large number of features. Due to these shortcomings in each model, the embedded model [9], was proposed to bridge the gap between the filter and wrapper models. First, it incorporates the statistical criteria, as filter model does, to select several candidate features subsets with a given cardinality. Second, it chooses the subset with the highest classification accuracy [4].

The embedded model usually achieves both comparable accuracy to the wrapper and comparable efficiency to the filter model. The embedded model performs feature selection in the learning time.In other words, it achieves model fitting and feature selection simultaneously.

Many researchers also paid attention to developing unsupervised feature selection. Unsupervised feature selection is a less constrained search problem without class labels, depending on clustering quality measures [2], and can eventuate many equally valid feature subsets. With high-dimensional data, it is unlikely to recover the relevant features without considering additional constraints. Another key difficulty is how to objectively measure the results of feature selection [2]. A comprehensive review about unsupervised feature selection can be found in [1]. Supervised feature selection assesses the relevance of features guided by the label information but a good selector needs enough labeled data, which is time consuming. While unsupervised feature selection works with unlabeled data but it is difficult to evaluate the relevance of features. It is common to have a data set with huge dimensionality but small labeled-sample size. High-dimensional data with small labeled samples permits too large a hypothesis space yet with too few constraints. The combination of the two data characteristics manifests a new research challenge. Under the assumption that labeled and unlabeled data are sampled from the same population generated by target concept, semi-supervised feature selection makes use of both labeled and unlabeled data to estimate feature relevance [7]. Feature weighting is thought of as a generalization of feature selection [69]. In feature selection, a feature is assigned a binary weight, where 1 means the feature is selected and 0 otherwise. However, feature weighting assigns a value, usually in the interval [0,1] or [-1,1], to each feature. The greater this value is, the more salient the feature will be. Most of feature weight algorithms assign a unified (global) weight to each feature over all instances. However, the relative importance, relevance and noise in the different dimensions may vary significantly with data locality. There are local feature selection algorithms where the local selection of features is done specific to a test instance, which is is common in lazy leaning algorithms such as kNN [22, 9]. The idea is that feature selection or weighting is done at classification time (rather than at training time), because knowledge of the test instance sharpens the ability to select features.

Typically, a feature selection method consists of four basic steps [4], namely, subset generation, subset evaluation, stopping criterion, and result validation.

In the first step, a candidate feature subset will be chosen based on a given search strategy, which is sent, in the second step, to be evaluated according to certain evaluation criterion. The subset that best fits the evaluation criterion will be chosen from all the candidates that have been evaluated after the stopping criterion is met. In the final step, the chosen subset will be validated using domain knowledge or a validation set.

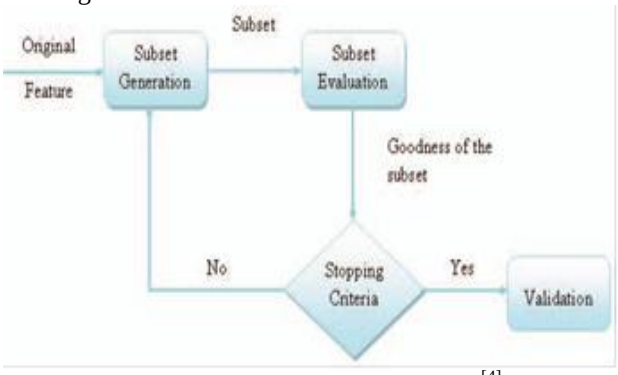

Fig. 4: Feature Selection Process [4]

A. Filter Algorithm:

Input: D(F0, F1, ..., Fn□1) // a training data set with N features
S0 // a subset from which to start the search
δ // a stopping criterion
Output: Sbest // an optimal subset

(1) begin
(2) initialize: Sbest = S0;
(3) ɤ best = eval(S0,D,M); // evaluate S0 by an independent measure M
(4) do begin
(5) S = generate(D); // generate a subset for evaluation
(6) ɤ = eval(S,D,M); // evaluate the current subset S by M
(7) if (ɤ is better than ɤ best)
(8) ɤ best = ɤ;
(9) S best = S;
(10) end until (δ is reached);
(11) return S best;
(12) end;

Fig. 5: A generalized filter algorithm [8]

B. Wrapper Algorithm:

Input: *D*(*F*0, *F*1, ..., *Fn*□1) // a training data set with *N* features
*S*0 // a subset from which to start the search
δ // a stopping criterion
Output: Sbest // an optimal subset

(1) begin
(2) initialize: Sbest = S0;
(3) ɤbest = eval(S0,D,A); // evaluate S0 by a mining algorithm A
(4) do begin
(5) S = generate(D); // generate a subset for evaluation
(6) ɤ = eval(S,D,A); // evaluate the current subset S by A
(7) if (ɤ is better than ɤ best)
(8) ɤ best = ɤ;

(9) Sbest = S;
(10) end until (δ is reached);
(11) return Sbest;
(12) end;

Fig. 6: A generalized wrapper algorithm [8]

VIII. MAX-REL FEATURE SELECTION

In feature selection, first we use InfoDist [8] to evaluate the distance between feature and class label. If a feature xi has less distance d(xi,C) with the class label C, we thought it is more relevant to the class label. We define a cutoff distance based on standard deviation. These features with distance larger than mean distance plus cutoff value are regarded as irrelevant and are removed. In experiment, we observe the effect to classification accuracy with respective to different cutoff values. Second, we use Pearson's correlation to evaluate the correlation between features. Two features with high correlation are redundant each other. We select the minimum redundancy features according to the correlation between the features. We select three different feature sets according to InfoDist distances and Pearson's correlations for our experiments. The three selection methods are described as follows:

A. Maximum Relevant Feature Set (MaxRel):

We use cutoff value to discard irrelevant features from the sorted InfoDist feature list. These features which have the smallest Pearson's correlation with respective to each feature in the remaining feature list are appended in the selection feature list with duplicates eliminated.

B. Minimum Redundancy Feature Set (MinRed):

We discard irrelevant features from the sorted InfoDist feature list using cutoff value as in maximum relevant feature set. The f e a t u r e which has the smallest Pearson's correlation with the listed feature is put immediate following the listed feature included in the selection list. The selected feature lists are primary based on less redundancy between features

1) InfoDist Calculation:

InfoDist is based on information theory [2]. The main concept of information theory is entropy, which measures the expected uncertainty or the amount of information provided by a certain event. The entropy of a random variable *X* is defined as follows:

$$H(X) = -\sum_{x} P(X = x) \log P(X = x) \tag{8.1}$$

Where P(X=x) is the prior probability of x.

Entropy *H(Y |X)* of a random variable *Y* given *X* is defined as follows:

$$H(Y|X) = -\sum_{x,y} P(x,y) \log P(y|x) \tag{8.2}$$

When the value of another random variable is known

The mutual information between two random variables *X* and *Y* is defined in the following:

$$\begin{aligned} I(X;Y) &= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)p(y)} \\ &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{8.3}$$

2) Pearson's Correlation Calculation:

If variables *X* and *Y* are continuous, the correlation is calculated by formula defined in the following:

$$r_{XY} = \frac{\sum xy}{n\sigma_x \sigma_y} \tag{8.4}$$

If X is a discrete feature with k values, and Y is a continuous feature. The correlation is calculated by formula defined in the following:

$$r_{XY} = \sum_{i=1}^{k} P(X = x_i) r_{X_{h_i} Y} \tag{8.5}$$

If variables X and Y are both discrete, the correlation is calculated by formula defined in the following:

$$r_{XY} = \sum_{i=1}^{k} \sum_{j=1}^{l} P(X = x_i, Y = y_j) r_{X_{h_i} Y} \tag{8.6}$$

Title	Proposes	Limitation	Future work
Classification for Multi-Relational Data Mining Using Bayesian Belief Network[5]	Proposed a Probabilistic Graphical Model, Bayesian Belief Network(BBN), based approach that considers not only attributes of the table but also the relation between tables	This approach has low classification accuracy	The implementation and the experimental results of the proposed approach will be carried out as future work.
A Novel Feature Selection Method Based on an Integrated Data Envelopment Analysis and Entropy Model[6]	Proposes both Data Envelopment Analysis which is a useful technique for determining the efficiency of decision-making units and Entropy method	It can only use for integer attribute. The proposed model can be used for ranking the features instead of selecting them	Suggest other researchers to use different MCDM methods such as TOPSIS and SAW integrated with other methods of weighting such as Expected Value method for selecting the features.
Competent MultiRelational Classifier using Filter based Feature Selection method on CrossMine	Explores CrossMine algorithm that use Tuple ID Propagation provides the way to directly search pattern from multiple tables with the use of virtual join	CrossMine algorithm is mixed ILP and relational database, improved the accuracy of traditional ILP and FOIL. But it is	With using the other type of feature selection method we can improve the accuracy

Algorithm[1]		still based on logical reasoning,	
Filter based Backward Elimination in Wrapper based PSO for Feature Selection in Classification[2]	Explores a new feature selection approach based on particle swarm optimisation (PSO) and a local search that mimics the typical backward elimination feature selection method.	it can also be observed that when the number of features reaches a certain small value, it may sacrifice the classification performance, like on the Vehicle and MultipleF datasets	In future also improve the performance of the PSO algorithm and also investigate a multiobjective feature selection approach
Comparison of wrapper and filtering approaches for corporate failure prediction[3]	Proposes five hybrid classifiers to tackle corporate failure prediction problem. filtering approach, genetic algorithm and particle swarm optimization techniques With knearest neighborhood (k-NN) to create our five classifies for our data set.	superiority of wrapper approach to filtering approach for this FDP data set.	In future the other hybrid techniques such as combination of PSO with logistic regression, GA with logistic regression can be applied to elicit the prediction accuracy for a real world corporate failure data set .
MR2 Based Feature Selection for Multi-Relational Naïve Bayesian Classifier[4]	Propose a feature selection method with multi-relational naïve Bayesian classifier. With these we can improve the classification accuracy and enhance comprehensibility of the models	The paper proposes the method is less scalable and give the slow performance	We may apply this technique to the more Relational dataset

Table 1: Analysis of Feature Selection Method

IX. Conclusion and Future Work

In feature selection technique there are different methods to apply on the different classifiers with different types of datasets. In filter method and wrapper there is efficient feature subset selection with better accuracy. In this type of approach it easily removes the redundant or irrelevant data, less time consuming and take less informational cost. It easily removes the redundant features like noise related feature etc. There are several problems in feature selection that can affect the preprocessing task. it require that more accuracy we get with efficient data. So there are various techniques used and also combination of different techniques gives the better performance. We can apply different feature selection techniques for different dataset.

X. Acknowledgment

This review is the part of M.E programme in Computer Engineering, Gujarat Technology University, Gujarat, India.

References

[1] "Competent Multi Relational Classifier using Filter based Feature Selection method on CrossMine Algorithm", IEEE, Nirma University International Conference On Engineering, India, 06-08december, 2012
[2] "Filter based Backward Elimination in Wrapper based PSO for Feature Selection in Classification", Institute of electrical and electronics engineers(IEEE), Beijing, China, July 6-11, 2014
[3] "Comparison of wrapper and filtering approaches for corporate failure prediction", Institute of electrical and electronics engineers(IEEE), Tehran, 2014
[4] "MR2 Based Feature Selection for Multi-Relational Naïve Bayesian Classifier", IJAR-CSIT Vol. 2 Issue 1 , 2013
[5] "Classification for Multi-Relational Data Mining Using Bayesian Belief Network", Springer International Publishing Switzerland, Dharmsinh Desai University, Nadiad, 2014
[6] "A Novel Feature Selection Method Based on an Integrated Data Envelopment Analysis and Entropy Model", Elsevier, 2nd International Conference on Information Technology and Quantitative Management, ITQM,2014
[7] "A Comparison of Filter and Wrapper Approaches with Data Mining Techniques for Categorical Variables Selection" International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 6, June 2014
[8] "MR-MNBC: MaxRel based Feature Selection for the Multi-Relational Naïve Bayesian Classifier", IEEE, Nirma University International Conference on Engineering (NUiCONE),2013
[9] "Feature Selection for Classification: A Review" Jiliang Tang, Salem Alelyani and Huan Liu
[10] "Feature Selection Methods And Algorithms" International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 5 May 2011

