Journal of Scientific & Industrial Research
Vol. 80, June 2021, pp. 537-542

An Optimized Approach for Feature Extraction in Multi-Relational Statistical Learning

Garima Bakshi[1], Rati Shukla[2], Vikash Yadav[3]*, Aman Dahiya[4], Rohit Anand[5], Nidhi Sindhwani[6] and Harinder Singh[7]

[1]Sushant University, Gurugram, India, [2]GIS Cell, MNNIT Prayagraj, Allahabad, India
[3]ABES Engineering College, Ghaziabad, Uttar Pradesh, India, [4]Maharaja Surajmal Institute of Technology, New Delhi, India
[5]G B Pant Engineering College, New Delhi, India, [6]Amity University, Noida, India,
[7]Sant Baba Attar Singh Khalsa College, Sandaur, Punjab, India

Received 05 December 2020; revised 12 February 2021; accepted 27 April 2021

Various features come from relational data often used to enhance the prediction of statistical models. The features increases as the feature space increases. We proposed a framework, which generates the features for feature selection using support vector machine with (1) augmentation of relational concepts using classification-type approach (2) various strategy to generate features. Classification are used to increase the productivity of feature space by adding new techniques used to create new features and lead to enhance the accuracy of the model. The feature generation in run-time lead to the building of models with higher accuracy despite generating features in advance. Our results in different applications of data mining in different relations are far better from existing results.

Keywords: Classification-type approach, Data mining, Feature space, Statistical model, Support Vector Machine

Introduction

Data analyst generally finds a very important topic in their research, to create a suitable reconstructing tool to convert the big data into more revealing and efficient form. Although particular specialised knowledge is required to design representation, generalised antecedent for learning data can also be helpful. The search for more efficient and useful representation of the data with the help of antecedent information motivates us. The objectives behind the data analysis as given by Breiman *et al.*[1]:

(1) To build stochastic systems which able to fit into the data and then draw some inferences from data generating method using structure of the systems.

(2) To predict the response of the system for the particular input variables.

In the recent years, the popularity and success of machine learning and deep learning methods in the field of text, image video and audio has given a chance for the researchers to give importance to the feature engineering.[2] These methods are efficient and effective when there is large numbers of computational tools and learning data are obtainable. The feature engineering wants more manual efforts in generating and selecting features, which is more complex and difficult. Learning most appropriate representation of feature is not insignificant job as it needs identifying manually the more complicated structure from data.[3] To overcome the problem of complication, you have to deal with all attributes and depend on the learning system to create important features.[4] However, this technique doesn't gives good result always. Most often working with such huge feature space is not effective as it increases the cost of computation because only some features are useful for the system.[5] It is also observed that the representation of such features is task-specific and suits accordingly.[6] So, there are many complex problems with this method. For the new problem, it is very difficult to decide which representations and its associated method should be chosen.[7] No existing methods are good one for the classification problem in real-world. The domain knowledge is useful for the generation of the features and thus, best expressed using feature constructed.[8] The existing methods are rigid, and do not allow user to change the representation.[9]

We present a relational statistical learning technique, generalised structural classification model (GSCM), for creating prediction model based on relational databases from different documents. In GSCM, features are created at run-time and tested for

*Author for Correspondence
E-mail: vikas.yadav.cs@gmail.com

better accuracy of the given model. This technique has many advantages as compared to different existing traditional techniques. The data from SQL queries are combined to give Boolean and quantitative data. The performance of the resulting model is better than the logical models. This will also show how the relational features combined with more features to improve the searching results accordingly.

Popescul *et al*. introduced a model to generate features in the relational DBMS using logistic regression with generating features from relational database.[10] They used the queries to retrieve the appropriate data from the database as search method. They also used richer join, grouping and argmax based queries in their search. They used their method to get the citations results in various publications.

The main features of the proposed research are as follows:

(1) We proposed novel feature generation and selection in relational database using Machine Learning called GSCM.

(2) We have applied natural join and SVM to increase the accuracy and therefore decrease the error.

(3) We are able to increase the performance of link prediction and venue prediction using CiteSeer dataset.

(4) We are also able to outperform the existing previous results.

Related Works

GSCM creates features are not similar to the previous generating features techniques.[11–14] In GSCM, a relational schema is generated based on queries results of database. Features are generated by combining different views using natural join then combine features are again filtered to get the most appropriate predictors for our model.

The relational schema is dynamically added containing some relations obtained by classification of data in the given tables. The classification is done according to keywords and authors name in the given published document and the relations between original data and the classified data obtained after classification based on keywords or authors name.[15]

All searches are based on the results of queries obtained by aggregation using natural join and according to predictive features generated so far. We will show that features generated with the help of queries required less computation than the previous method.

GSCM combines two basic techniques i.e. supervised model capable of outperform logical models and better technique to extract features from the multi-relational views. Classification models are more accurate than the traditional logical methods. This difference is achieved when bulk features are added to models, many of them shows some good results because word are taken as features. Classification models also used some very good feature extraction techniques such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), Gabor Filters, Layered Approach, Geometric Features, Moments Invariants, Zernike Moments and Local Binary Patterns to overcome over fitting problem.[16]

The important components of GSCM system are described in Fig. 1. Two basic components: feature generation and classification model – are bonded in the single loop. Samples are taken from the population of the relational database. Queries are executed to get the various views, which are then aggregated by the aggregation function such as natural join. Features are generated from the combined views of the relational model. Further, classification model are selected and assessment of its performance is done on the basis of accuracy. Based on the performance of the classification model some more features are added to enhance the performance of the model. This is basically recursive model, used to evaluate its own performance.

GSCM has various key properties which makes it better than regression models and ILP:

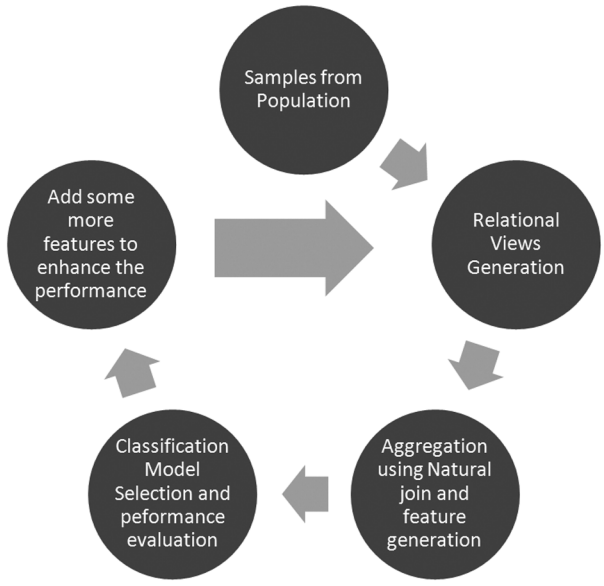

Fig. 1 — Learning Process

- The use of classification model instead of regression and logic permits us to use more effective summaries of the keywords used for searching in any document.
- The use of classification model for generating new features from the set of relations. Classification gives more useful model for relational data and produce better representation and scalability. For example, classification of the documents based on the given keyword can be easily separated.
- The use of relational database management system and SQL produce more realistic results as compared to PROLOG. Most of the data in real world depends on RDBMS due to its scalability and optimisation techniques, having more information and metadata.
- Binding up of feature generation and classification model in the single loop gives more efficient result as compared to pre-feature generation technique. Features are not generated in advance but generated in run time which allows us to give more attention on useful features. They can be generated one by one and keeping only few selected features.

The results are verified on two sets of task applied on the data in the CiteSeer, online database of different journals. It contains different paper title, abstracts, keywords and journal names. We used CiteSeer as a relational database for our experiments. The results based on keywords and last name of authors are analysed using machine learning.

Relations for Classification

GSCM uses classification to produce new relations and incorporated them into schema for statistical learning technique. Entities and relationship produced from the classification increase the productivity of the feature space. These features are added to the database schema which are then utilized to enhance the performance of the queries (Fig. 2). For example, in CiteSeer papers are classified based on words or keywords and on authors with the journal name. The original database is used to describe which entities to be taken and relations between these entities for classification of the results. For example, database can be classified using keywords, authors and journal name. Some important features are selected among so many features, kept in the classification model, based on the statistical selection technique. Classification enhances the accuracy as well as generation of efficient predictive features.[17,18]

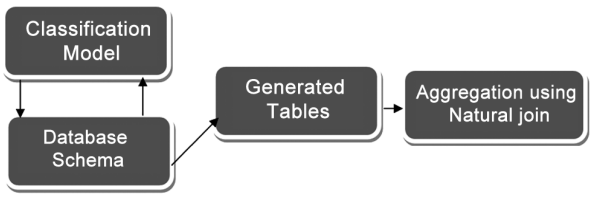

Fig. 2 — Feature generation from database schema

The aggregation used in the model works differently from cluster and ILP; produce some new features from the relational database.[11] Aggregation is used to represent the results from the relational database in the tabular form, and some logical queries in the scalar form. For example, average number of keywords used in the research paper and counting the citations of each paper in the journal. Aggregation when incorporated with classification is very efficient to create new features and then this model used repeatedly to create new queries and tables to enhance the performance of the model.

Run-time Feature Generation

GSCM also supports feature generation and selection at run-time. The most challenging and computational part of GSCM is feature generation. The generation of huge features takes lot of CPU time to get the results of SQL queries, but our model generates more efficient features with fewer calculations at run-time. The problem of over fitting should not be ignored even if we generate the features and also test them to include in the model is determined at run-time. The aggregation using natural join reduces the calculations.

Query results are categorised into various streams based on the different expressions taken by the users. Size of each feature is different for example; number of keywords is greater than the number of journal. It is very easy to classify selected features into different categories. The features which utilize most of the other features are chosen to get the next generated feature. The query having most dominating results among all queries should be taken first.

Proposed Methodology

GSCM binds two basic concepts: feature generation from relational database, feature selection using classification model. We developed an optimised algorithm for GSCM. In GSCM, views are aggregated using Natural Join and classification using SVM. The efficient algorithm depends upon the CPU and accuracy of the model. These two factors decide when to stop the algorithm.

We use SQL, due to its connectivity and efficiency with database engines.[19] We use the given schema in this paper:

Keyword (Documents, count, Keyword)

Authors Publications (Documents, Journal name)

Citations (SourceDoc, Target Doc)

Coauthors (Documents, Coauthors)

The domains are different from the SQL. Individual queries produce tables for each query. Further, they are aggregated using natural join. Aggregate functions are used with subscript with the variables. For example, the sum of counting of word "Human" in the database where document is written by author or co-author is given by:

Results (R) = Sum (AuthorsPublications (D, N) ⋈ Keyword (D, count, Human))

The RHS of the above expression is used to generate features to test our classification model.

SELECT DISTINCT *

FROM Author Publications R1

WHERE R1.D='N' AND R1.D='Human'

The above query produces the results that counting of "keyword" in the document for the given "author".

Experiments and Results

We evaluate GSCM using CiteSeer database, a relational online journal of computer science. This database includes author name, title of the paper, journal name, citation information etc. We used the same database as in the work by Popescul *et al*.[10] Further, we try to compare the results of them with the results of our GSCM model.

There are 2122 journal unique journals and conferences, 32640 coauthors are present,

123420 are the total words and 27820 are the total citations.[20] The relations and their instances are shown in Table 1.

We produce results of combination of two tasks: Author total publications and in the publication count the frequency of given word. The two tasks are Authors Publications (Documents, Journal name) and Keyword (Documents, count, Keyword). In both tasks the search space is based on the given relations such that quantity of authors, journal and conference, frequency of words etc. and response are Boolean in nature.

Above three relations are many-to-many relations and one is one-to-one relation. Model is learned using sequential feature selection i.e. each feature is generated, it is added to the model if accuracy improves otherwise removed from the model. We use n-fold cross validation rule to enhance the model using SVM classification rule. All observations are divided equally into n sets, n-1 set is used to train the model and each set is used to test the model. In the prediction of venue (Journalname), there are 12000 observations. Out of 12000, 6000 are the positive taken out from the relation <Document, Journalname> and 6000 are negative samples taken out randomly from the remaining documents. The size of the Author publications is decreases by 6000 after removing positive samples. In the prediction of the link, the 6000 samples are taken for the experiment. Out of 6000 samples, 3000 are the positive samples taken from the <sourcedoc, targetdoc> and remaining 3000 are the negative samples taken randomly from the database. The number of samples of citations decreases by 3000 due to positive samples.

A total of 3500 features are used to train the classification model, equal to the total features considered by Popescul *et al*.[10] The features are numeric in nature because it is easy to compare the difference in accuracy of both model i.e. our model and model of Popescul *et al*.

The accuracy of our model with respect to features considered from 0 to 3500 in both prediction of venue as well as link are presented in Table 2. The

Table 1 — Total relations and instances

Relations	Size
Authors Publications (Documents, Journalname)	2122
Citations (SourceDoc, TargetDoc)	27820
Coathors (Documents, Coauthors)	32640
Keyword (Documents, Count, Keyword)	123420

Table 2 — Accuracies with number of features considered in venue prediction and link prediction

Features considered	0	250	500	750	1000	1250	1500	1750	2000	2250	2500	2750	3000	3250	3500
Accuracies in venue prediction	0	74.1	76.2	78.1	81.3	80.2	80.9	86.5	85.9	86.1	86.2	86.2	86.2	86.4	86.5
Accuracies in link prediction	0	83.8	92.1	91.4	92.2	91.9	91.3	90.4	92.3	92.1	92.4	92.3	92.5	92.4	92.5

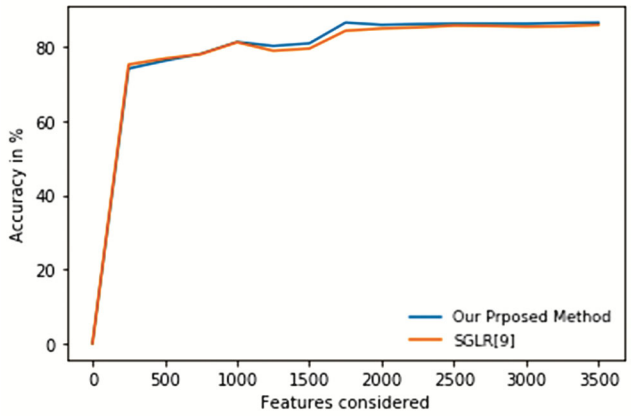

Fig. 3 — Comparison between SGLR method with proposed method in Venue Prediction

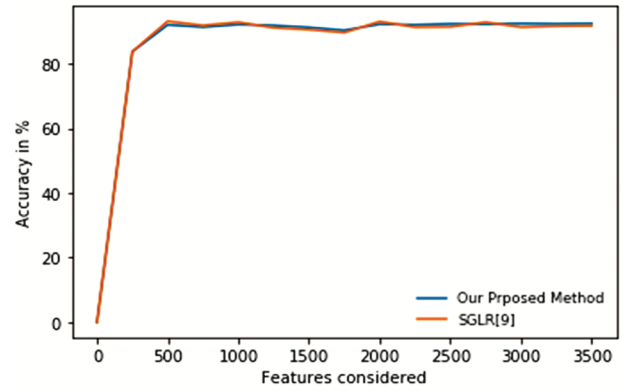

Fig. 4 — Comparison between SGLR method with proposed method in Link Prediction

difference between the plot of average accuracy of our model for the features considered with the Popescul's model for the venue and link prediction is shown in Fig. 3 and Fig. 4. The plot shows that the accuracy of prediction changes with addition of more features at equal intervals and also shows better performance of our proposed model with Popescul's model in terms of accuracy. The average of our accuracy found to be 82.1% and 91.3% for the prediction of venue and link prediction, which is better than the Popescul's accuracy. The difference between accuracy in the prediction of venue and link are presented in Fig. 5 and Fig. 6 respectively.

The classification relations above have higher accuracy as compared to cluster relations. Another advantage of working in SVM classification model is that it is cheaper than the cluster relations because it filter all the data from the database. This also reduced the cost of computation for producing features from the relational database.

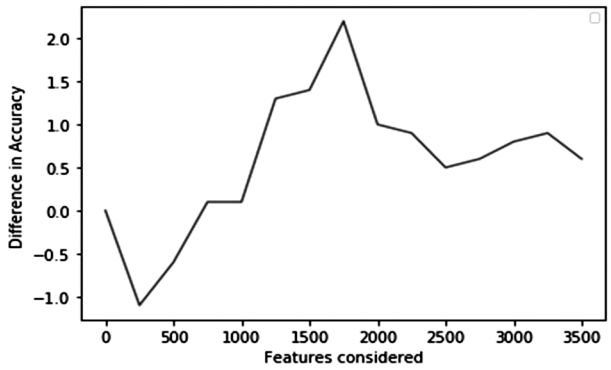

Fig. 5 — Difference between accuracy between Proposed Method with SGLR method in Venue Prediction

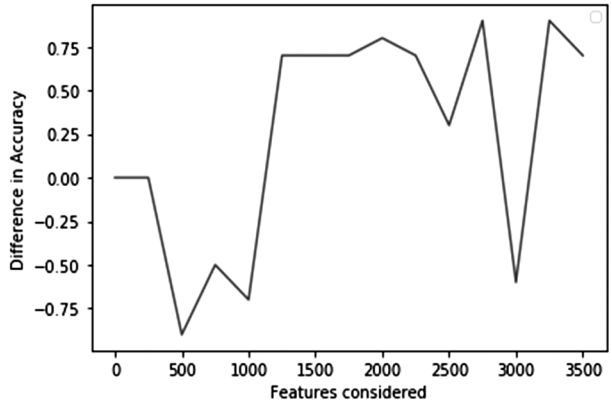

Fig. 6 — Difference in accuracy between Proposed Method with SGLR method in Link Prediction

Conclusions and Future Works

We introduced Generalised Structural Classification Model (GSCM), which uses natural join and SVM for the analysis of CiteSeer database. GSCM combines the power of natural join and SVM with automatically generation of features from the relational database. Existing models generate features comparably slow, derived rigorous procedure produced from the regression and ILP based models. GSCM is applicable to large database which are noisier, complex and distributed in nature, efficient feature generation from the large database, query optimisation makes it more efficient, robust and accurate model.

We also showed how SVM can be used to derive new important features from the feature space in the relational statistical learning. SVM improves accuracy with the help of dimensionality reduction. Entities build from our model increase the expressivity of the given feature space.

We also showed that feature generation in advance is less reliable than the feature generation at run time

which requires fewer calculations. Our model is also able to reduce computational cost to some extent. Our model will very helpful to the modelling of hyperlinks, social networks, bioinformatics, biosciences and some control over statistical database. In the future, some more machine learning techniques will be used in the given model to improve its accuracy such as Logistic Regression, Naïve Bayes, KNN etc.

References

1 Breiman L, Statistical modeling: The two cultures (with comments and a rejoinder by the author), *Statist Sci*, **16(3)** (2001) 199–231, https://doi.org/10.1214/ss/1009213726

2 Bosch S V D, Automatic Feature Generation and Selection in Predictive Analytics Solutions, Master's Thesis, Faculty of Science, Radboud University, Nijmegen, The Netherlands, (2017).

3 Ding Y, Zhou K & Bi W, Feature selection based on hybridization of genetic algorithm and competitive swarm optimizer, *Soft Comput*, **24** (2020) 11663–11672, https://doi.org/10.1007/s00500-019-04628-6.

4 Akbal E & Tuncer T, FusedTSNet: An automated nocturnal sleep sound classification method based on a fused textural and statistical feature generation network, *Appl Acoust*, **171** (2021), https://doi.org/10.1016/j.apacoust.2020.107559.

5 Tuncer T, Dogan S & Subasi A, A new fractal pattern feature generation function based emotion recognition method using EEG, *Chaos Soliton Fractal*, **144 (C)** (2021), https://doi.org/10.1016/j.chaos.2021.110671.

6 Sreevani C A & Murthy B C, Generation of compound features based on feature interaction for classification, *Expert Syst Appl*, **108(15)** (2018) 61–73.

7 A Pillai, R Soundrapandiyan & S Satapathy, Local diagonal extrema number pattern: A new feature descriptor for face recognition, *Future Gener Comput Syst*, **81** (2017) 297–306.

8 Bharot N, Verma P, Sharma S & Suraparaju V, Distributed denial-of-service attack detection and mitigation using feature selection and intensive care request processing unit, *Arab J Sci Eng*, **43(2)** (2018) 959–967.

9 Gallitz O, Candido O D, Botsch M & Utschick W, Interpretable feature generation using deep neural networks and its application to lane change detection, *IEEE Intel Transport Syst Conf (ITSC)*, Auckland, New Zealand, 2019 3405–3411.

10 Popescul A, Ungar L H, Lawrence S & Pennock D, Statistical relational learning for document mining, *Proc IEEE Int Conf Data Mining*, (2003) 275–282.

11 Perlich C & Provost F, Aggregation-based feature invention and relational concept classes, *Proc Int Conf Know Discov Data Mining*, 2003, 167–176.

12 Laer W V & Raedt L D, How to upgrade propositional learners to first order logic: A case study, *Relational Data Mining*, Springer-Verlag, Berlin, 2001, 235–261.

13 Dehaspe L, Maximum entropy modelling with clausal constraints, *Proc Int Conf Induct Logic Program*, **1297** (1997) 109–124.

14 Roth D & Yih W, Relational learning via propositional algorithms: An information extraction case study, *Proc Int Joint Conf Artif Intell*, **2** (2001) 1257–1263.

15 Shi H, Li H, Zhang D, Cheng C & Cao X, An efficient feature generation approach based on deep learning and feature selection techniques for traffic classification" *Computer Networks*, **132** (2018) 81–98.

16 Rahul M, Shukla R, Yadav D K & Yadav V, Zernike moments based facial expression recognition using two staged hidden markov model, *Adv Intell Syst Comput*, **924** (2019) 661–670.

17 Yadav V, Rahul M & Shukla R, A New Improved Approach for Feature Generation and Selection in Multi-Relational statistical modeling using machine learning, *J Sci Ind Res*, **79(12)** (2020) 1095–1100.

18 Singh S & Yadav V, Face recognition using HOG feature extraction and SVM classifier, Int J Emerg Trends Eng Res, World Academy of Research in Science and Engineering Publication, **8(9)** (2020) 6437–6440, DOI: 10.30534/ijeter/2020/244892020282.

19 Ceri S, Gottlob G & Tanca L, Logic programming and databases", Springer-Verlag, Berlin, **1,** 1990, 284–294.

20 http://dblp.uni-trier.de/