|  |
| --- |
| [SIGMOD2015](http://www.sigmod2015.org/" \t "_blank) **SIGMOD2015** May 31- June 4, 2015, Melbourne, Australia |
|  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Meta-Reviews For Paper**   |  |  | | --- | --- | | **Track** | Research 2nd Submission | | **Paper ID** | 256 | | **Title** | MRLBase: Multi-Relational Learning with Multi-Relational Databases |  |  |  | | --- | --- | | **Masked Meta-Reviewer ID:** | Meta\_Reviewer\_1 | | **Meta-Reviews:** |  |  |  |  | | --- | --- | | **Question** |  | | Summary of reviews and discussion | The paper shows how to train a Bayesian network in a relational database using SQL. The paper presents a competent approach to solve an important problem. However, the incremental contribution of this paper given the extensive prior work on the topic is not substantial enough for SIGMOD. | | Recommendation | Reject | |

|  |
| --- |
| [SIGMOD2015](http://www.sigmod2015.org/) **SIGMOD2015** May 31- June 4, 2015, Melbourne, Australia |
|  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Reviews For Paper**   |  |  | | --- | --- | | **Track** | Research 2nd Submission | | **Paper ID** | 256 | | **Title** | MRLBase: Multi-Relational Learning with Multi-Relational Databases |  |  |  | | --- | --- | | **Masked Reviewer ID:** | Assigned\_Reviewer\_1 | | **Review:** |  |  |  |  | | --- | --- | | **Question** |  | | Overall Rating | Reject | | Summary of the paper (what is specifically being proposed, and in what context) and a brief justification for your overall recommendation; one paragraph | The authors present techniques for doing machine learning by leveraging in a major way the machinery of relational DBMS. | | Three (or more!) strong points about the paper. Please be precise and explicit; clearly explain the value and nature of the contribution. | S1. The topic is very important.   S2. The idea of doing as much as possible within a DBMS is very appealing.   S3. The paper clearly has good ideas. | | Three (or more!) weaknesses of the paper. Please indicate clearly whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution. Write it so the authors understand what are seen as the negatives. | W1. The writing is such that it is not possible to learn from this paper how to do learning using relational DBMS. Major flow and explanation-quality problems abound in the paper, which makes the paper unusable in its current form.   W2. There are no experimental comparisons to the state of the art, and no clear explanations of why such comparisons may not be feasible.   W3. The paper is not self contained: One has to read a lot outside the paper to be able to understand the exact nature of the contributions. | | Is the paper relevant for SIGMOD? | Yes | | Significance | The paper improves on existing work | | Technical depth and quality of content | Syntactically complete but with limited contribution | | Validation - experiments and proofs | Unclear/obscure, hard to determine what is going on and what has been validated | | Presentation | Sub-standard: needs a heavy rewrite | | Discussion of related work - recall that the new page limits allow for material outside the 12pp including references, hence we expect good coverage of related work | Some description of related work, but could provide better perspective | | Detailed Evaluation (contribution, pros/cons, errors); please number each point | Please see the strong and weak points. | | If revision is required, please list specific revisions you seek from the authors | A very significant rewrite of the paper is needed. |  |  |  | | --- | --- | | **Masked Reviewer ID:** | Assigned\_Reviewer\_2 | | **Review:** |  |  |  |  | | --- | --- | | **Question** |  | | Overall Rating | Accept | | Summary of the paper (what is specifically being proposed, and in what context) and a brief justification for your overall recommendation; one paragraph | This paper proposes a multi-relational learning framework, MRLBase, on top of DBMS. By formalizing the learning operators as database queries, MRLBase can support various learning jobs, such as bayes network. The beauty of MRLBase is that it does not introduce any new overhead of using MRLBase. All functions are implemented using the DBMS. | | Three (or more!) strong points about the paper. Please be precise and explicit; clearly explain the value and nature of the contribution. | 1. It is an interesting work to design a learning framework on top of the DBMS.  2. The design is elegant without too much intrusion to the DBMS.   3. Bayes network on multiple relations can be efficiently trained via MRLBase.  4. The presentation is good and easy to follow. | | Three (or more!) weaknesses of the paper. Please indicate clearly whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution. Write it so the authors understand what are seen as the negatives. | 1. There is few discussions on the performance issues. The basic idea behind MRLBase is similar to the data cube. All require to pre-compute the group-by aggregations (e.g., counts). Joining multiple large tables are definately costly.  2. As pointed by previous work in multi-relation mining area, joining tables together will result in a different semantics from mining rules from each table individually and then combining them. The paper should give a clarification.  3. The paper does not show how MRLBase can be applied to other machine learning jobs. | | Is the paper relevant for SIGMOD? | Yes | | Significance | The paper improves on existing work | | Technical depth and quality of content | Solid work | | Validation - experiments and proofs | OK, but do not cover all of the claims | | Presentation | Excellent: careful, logical, elegant, understandable | | Discussion of related work - recall that the new page limits allow for material outside the 12pp including references, hence we expect good coverage of related work | Clear explanation of the state of the art and how this paper relates | | Detailed Evaluation (contribution, pros/cons, errors); please number each point | 1. I like the idea, which is similar to the Madlib proposed by Berkeley. All rely on the SQL (and its backend database engine) to process complex jobs. In fact, SQL is a very expressive language. It can do more than we expect.  2. The paper formalizes the learning problems into some operators which can be implemented as SQL statements. This is good, as we do not need modify the DBMS itself.  3. However, there are still some problems that are not well addressed. First, there is no optimization for the performance. So far, the paper just simplifies joins all involved tables and performs the counting for different groups. This approach is similar to build a data cube which is very expensive for large datasets. Second, it is unknown whether the framework can support other learning jobs, except the Bayes network, since it mainly depends on the counts.  4. No discussion on update. I believe your contingency table is built offline although it is clearly specified. What if new data are inserted and how would you update that contingency table? | | If revision is required, please list specific revisions you seek from the authors | Please answer the questions in weak point 2 and 3.  Perform some experiments to show the performance. |  |  |  | | --- | --- | | **Masked Reviewer ID:** | Assigned\_Reviewer\_3 | | **Review:** |  |  |  |  | | --- | --- | | **Question** |  | | Overall Rating | Reject | | Summary of the paper (what is specifically being proposed, and in what context) and a brief justification for your overall recommendation; one paragraph | This paper studies the problem of implementing multi-relational learning algorithms using relational databases to achieve better scalability.  In particular, the authors study three specific problems:  1. How do you represent random variables in terms of relations 2. How do you learn a Bayesian Network from multiple relations by using SQL to compute sufficient statistics 3. How do you perform inference on the learned Bayesian Network. | | Three (or more!) strong points about the paper. Please be precise and explicit; clearly explain the value and nature of the contribution. | 1. Important problem 2. Approach seems to speed up training time relative to some other techniques | | Three (or more!) weaknesses of the paper. Please indicate clearly whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution. Write it so the authors understand what are seen as the negatives. | 1. Low on technical meat 2. Missing some related work 3. Hard to understand due to poor paper organization | | Is the paper relevant for SIGMOD? | Yes | | Significance | The paper will have no impact | | Technical depth and quality of content | Syntactically complete but with limited contribution | | Validation - experiments and proofs | Unclear/obscure, hard to determine what is going on and what has been validated | | Presentation | Sub-standard: needs a heavy rewrite | | Discussion of related work - recall that the new page limits allow for material outside the 12pp including references, hence we expect good coverage of related work | Some description of related work, but could provide better perspective | | Detailed Evaluation (contribution, pros/cons, errors); please number each point | This paper studies the problem of implementing multi-relational learning algorithms using relational databases to achieve better scalability. In particular, the authors study three specific problems:  1. How do you represent random variables in terms of relations 2. How do you learn a Bayesian Network from multiple relations by using SQL to compute sufficient statistics 3. How do you perform inference on the learned Bayesian Network.  While the problem is a natural (and presumably useful) one, the technical contributions of this paper, relative to the prior work either considering the single table case, or on studying other representations (e.g., MLNs instead of Bayes nets) is not very substantial.   Most of the novelty in this paper is in the translation of operations performed on the Bayes net into SQL, and the representing and storing the statistics as relations, which is not very surprising or challenging given the vast quantity of prior work.    Another few related papers:  The Missing Piece in Complex Analytics: Low Latency, Scalable Model Management and Serving with Velox: http://arxiv.org/abs/1409.3809  Felix: Scaling Inference for Markov Logic with an Operator-based Approach  Incrementally Maintaining Classification using an RDBMS  GeoDeepDive: Statistical Inference using Familiar Data-Processing Languages    Further, the paper is not written in a style that is easy to read: there are no examples until the 4th section, and most of the techniques are described "in the abstract". I encourage the authors to bring the examples up into Section 1 itself, to better situate what they are trying to accomplish.   In terms of improving the technical depth of the paper, I encourage the authors to consider the following directions: 1. How to learn a Bayesian Network from incomplete datasets? -- Based on my understanding, if part of the attribute values in some tuples are missing, the technique described in the paper would no longer work. 2. How to efficiently perform maximum likelihood inference over multiple random variables? -- We can imagine that when new data arrives, we would like to find the maximum likelihood estimation of unknown attributes given their connections with existing data. In the case when the unknown attributes are correlated with each other, we may need to perform joint inference over all variables. I understand that the techniques in the paper can be used to compute the joint distribution of multiple random variables, but that might be an overkill if we are only interested in maximum likelihood estimation. | | If revision is required, please list specific revisions you seek from the authors | -- | |  |