Relationship between disparities of produced values from Aggregate Functions and Bayes Factor

By
Dr. Saleh Al-Qaraawy Dr. Jane Jaleel Stephan

And Miss Lamia Abed Noor

Abstract

The aggregate functions are used in extraction features from groups of data. These functions are varied in nature of produced features that are represented by aggregate values. Our work in this paper is interested with discussing the existence of any relationship between the disparity of produced aggregate values and the bayes factor of the Probabilistic Relational Models based aggregate. Through this work the practical experiment was performed using experiment data in order to support the paper conclusions.

Keywords: Probabilistic Relational Models, Bayes Factor,Aggregation.

1. Introduction

Probabilistic Relational Models(PRMs) are models that are used in mining relational data base. However the descriptive attributes in tables are introduced as random variables that will be seek for the dependencies between them in order to be in target model, the dependencies may be between the attributes in one or in multiple tables which are associated by relationship addressed with different cardinality and this requires for special treatment. If the relationship is addressed with cardinality”one-to one” then the attributes will be as in one table. While the relationship, is addressed with “one-to-many” needs to convert the many value to one and this can be done with aggregation tool. The multiple values can be concluded in single data that represent some feature for the concluded data according to the aggregate function. The concluded data that are produced may not represent the original data and they may be dispirited from the expected data and this varied from function to another. As a result the models which include these data will be affected. In order to evaluate the constructed models, Bayes Factor is one of these tests that are used in model selection.

2. Probabilistic Relational Models

The data generated and collected from various activities reflect the life, which contains different objects that interact to find the suitable tools, which organize them in response to their features. Relational database is a tool for management of the data as separate tables that contains individual objects and at the same time introduce explicit relationships between them.

Probabilistic relational models (PRMs) are tools for mining the relational database in order to extract the regular features of interesting data and formulate them in suitable forms that are represented by probabilistic models. PRMs are extended traditional models with flat table. PRMs, in contrast to the traditional models devoted to the attributes of a single table, exploit the structure of relational database in addition to the attributes that they contain.

Probabilistic Relational Models (PRMs) are a recent development[1] that extends the standard attribute-based Bayesian network representation to incorporate a much richer relational structure. A probabilistic relational model specifies a template for probability distribution over a database. The template allows a generic dependence between attributes of the tables to be represented for a class of objects, which is then initiated for particular sets of entities and relations.

PRM consist of set of pairs (Child|Parents), each one is called local structure. The random variables are introduced in the model as child or parent. The score of model is the sum of scores of local structures, so it can be maximized through maximizing the local structures.

3. Probabilistic Relational Model Based Aggregation

The relational model in relational database introduced has rich information about associations. The relationships that connect between different tables are identified as one-many or many- many, more features can be added to be exploited in the dependent network using a suitable function that is called aggregate function. However, the aggregate functions are varied; the features provided will be different. They are used in order to introduce a new value which describes multiple values.

Course Table

Course code	Course type	Lecturer code
C_0	Software Engineering	L_0
C_1	Programming	L_0
C_2	Programming	L_0
C_3	Programming	L_0
C_4	Programming	L_1
C_5	Data mining	L_1
C_6	Software engineering	L_2
C_7	Operating System	L_3
C_8	Communication	L_3
C_9	Programming	L_2
C-10	Data Mining	L_1
C_11	Programming	L_3
C_12	Software Engineering	L_2

Lecture Table

Lecturer_code	Lecture_ name
L_0	X
L_1	Y
L_2	Z
L_3	M

Figure(1) university data base

Aggregate functions are functions on sets of records. Given a set of records and some instance of an aggregate function, a feature of the set which summarizes the set is computed. Then, aggregates are used to characterize the structural information that is stored in tables and associations between them. Consider two tables P and Q, linked by a one-to-many association A. For each record in P, there are multiple records in Q, and A defines a grouping over Q. An aggregate function can now be used to describe each group by a single value. Because a single value is specified for every record in P, we can think of the aggregate as a virtual attribute of P[2]. As result, P will have additional attribute and can be searched for existence of its interaction with other attributes in P.

The aggregate functions are classified according to the kind of the aggregated values, the continuous values; MAX, MIN, MEAN and so on. While the discrete values, “MODE” is the conventional function, which be used in order to summarize multiple discrete values in one through selection of the state that have higher frequency. Figure(1) describes an example of relational database that are specified to university, each record in table “Lecture” may connect to one or more records in table “Course”. Therefore the values of attribute “course_type” in Course must be converted from multiple values, which are related with one value in “Lecturer” table to single value in order to enable it with face to attributes in “Lecturer” table and introduce the summarized values in virtual attribute that can be imagined as additional attribute in relation “Lecture” as shown in Figure(2).

Lecturer Table

Lecturer_code	Lecturer_name	Aggregate “Course_type”
L_0	X	Programming
L_1	Y	Data Mining
L_2	Z	Software Engineering
L_3	M	Programming

Figure(2) Lecturer Table with “virtual attribute” hold aggregate values result from attribute”course_type” in “Course” table

The new virtual attribute can be dealt as a random variable and introduced in the hypothesis space in order to find the intended model.

4. Bayes’ Factor(BF):

It is a tool for evaluation the candidate models; which model must be preferred i.e. for two models M1, M2:

$$BF_{12}=\frac{P(M1|D)}{P(M2|D)}=\frac{P(M1)P(D|M1)}{P(M2)P(D|M2)} \qquad (1)$$

The first ratio (P(M1)/P(M2)) measure how much prior beliefs favored M1 on M2, this is depended on the available prior knowledge[3]. While the second ratio (P(D|M1)/P(D|M2)) expressed how well the observed data were predicated by one model compared to another.

Where P(D|Mi) is the probability of the observed data across all parameter values. The term corresponding to an average of likelihoods those are under a prior distribution of the parameters[4].

Bayes Factor ratio is interpreted as a summary of the evidence for M1 against M2 according to the measures that is shown in Table (1).

Table(1) Bayes Factor Evidence[4]

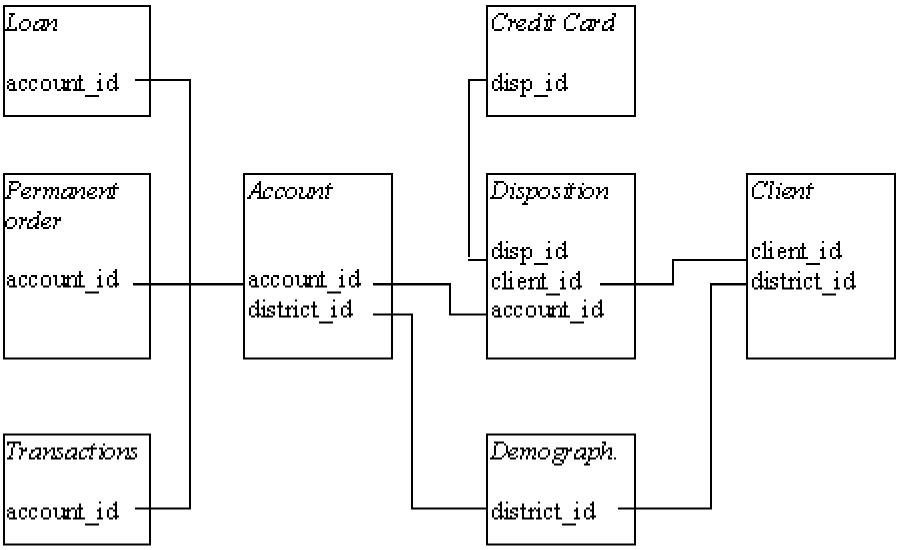
BF_{12}	Evidence for M_1
<1	Negative (Support M_2)
1 to 3	Not worth more than a bare mention
3 to 12	Positive
12 to 150	Strong
>150	Decisive

5. Practical Work

The aim of the practical work is to show out the effect of disparities of aggregated attribute on the target model through the measure of model evaluate, Bayes Factor. This can be accomplished when using different aggregate functions which produce different values and as a result will have different disparities. Two aggregate functions were applied in this work; one of these functions was addressed by “conventional”, adopts the value with high frequency in bag of instances, while other was referred to “proposed” which emerges value with scarce and importance frequency. Then construct two groups of model; a group of models that formulate aggregate values that would be produced from conventional function, and other group of models which contain aggregate values from proposed function.

5.1. Expirement Data Description

The experiment data we use are from PKDD'99 Discovery Challenge “Guide to the Financial Data Set” and are given free for research purposes[5] as shown in Figure(1).



Figure(1) Relational Structure Schema Of Experimented data

This schema has seven tables that are associated with different cardinality. Table "Transaction" which contains relation **transaction** 1056320 instances and table "Account" which contains 4500 instances, were associated with cardinality with "many-to-one" and this is suitable for the practical work in this paper. The descriptive attributes of "Transaction" that were tested are "Date", "Type", "Operation", "k-symbol", "Bank_to". In opposite table "Account", there are two descriptive attributes; "Date" and "Frequency".

5.2 Disparity Measure

Disparity is a tool that can be used in order to evaluate each of aggregate functions, the disparity of the resulting states is computed for each attribute, where it refers to the difference sums of the absolute distances(dis) of aggregate value frequencies($freq_{agg}$) and the expected value frequencies($freq_{exp}$) as :

$$dis= \sum_{i \in Val(X)} abs(freq_{agg}(i) - freq_{\exp}(i)) \qquad (2)$$

The expected value frequencies can be obtained from the frequencies that are exists in the source data as: -

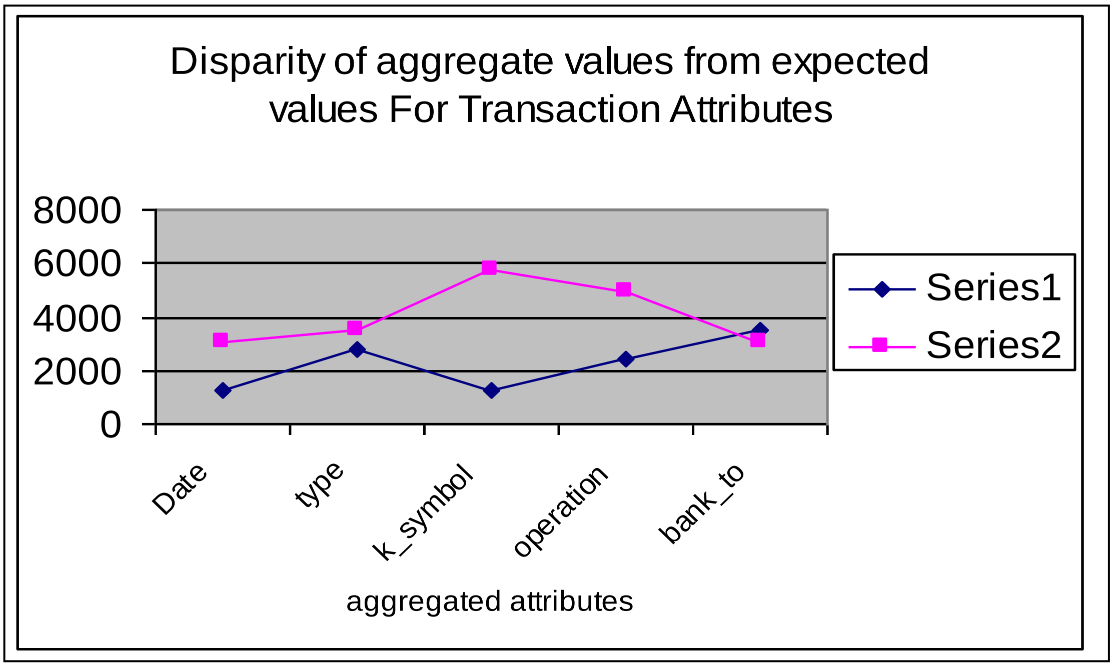
Let intended aggregate attribute in table(X) and aims to aggregated according to associate tables (Y)

$$freq_{exp}(i)=(freq(i)/size(X))* size(Y) \qquad (3)$$

According to our experimented data set where “Transaction” relation indicates to relation(X) and account relation indicates to (Y).

$\text{freq}_{\text{exp}}(i)=(\text{freq}(i)/\ 1056320)*4500$

Applying this test on the aggregate attributes in relation “Transaction” for both approaches, we see the results are shown in Figure(2). Thdisparities of aggregated values for most of tested attributes (Date, Type, Operation, and K_symbol) in applying the proposed method are lower than the aggregated values that are resulted from applying conventional approach.



Figure(2) Disparity Measure of aggregate value with expected ones. Series 1 refers to disparity of aggregate values resulted from proposed approach while series 2 indicates to disparity of aggregate values produced by applying conventional

5.3 Construction Models

After the aggregate values were produced from specified attributes according to different functions, they were presented as random variable and then they were tested in cooperated with the attributes in opposed table that is “Account” in order to locate the pair of dependent random variables that can be formulate as (Child|Parent) and then may be in candidate model. The test of dependence was applied on the random variable and only the pairs of random variable that are shown in Table(2) were succeeded.

Table(2) candidate models from local structures(Child|Parent)

No.	Models
1	Account_Date| Transactions_Date
2	Transactions_Type|Account_Date
3	Account_Date|Transactions_Operation
4	Account_frequency|Transactions_Date
5	Transactions_Type|Account_frequency
6	Account_frequency|Transactions_Operation

As a result, there are 6 pair models, each pair of models represents two models; model contains aggregate attribute based on conventional function and other contains aggregate attribute based on proposed functions. So the comparison that may be accomplished is executed between these pair of models according to function that was executed.

5.4. Scoring Bayes Factor for candidate models

The produced candidate models for previous step were tested against Bayes Factor for two groups of models; it was computed for each pair of models. The result of scoring the bayes factor are shown in Table(3). It is consist of value that compare two model based on aggregate function(conventional|proposed). As shown in Table(3), the results of bayes scores indicate to prefer the models that are based on conventional aggregate function for most of candidate models. According to Table(1) which evaluates the scores of Bayes Factor, some of models win the degree positive for model(based conventional function) against mode(based proposed function) such as score for model (no.2) in Table(3)[6].

Table(3) Bayes Factor for comaprision pairs models constructing from(conventional and proposed aggregate functions)

No.	Candidate Models	Bayes Factor (conventional|proposed)
1	Account_Date|Transactions_Date	0.2728235 0.20072
2	Transactions_Type|Account_Date	0.7794089 0.1775084
3	Account_Date|Transactions_Operation	0.1866556 0.0926731
4	Account_frequency|Transactions_Date	0.8880269 0.8544284
5	Transactions_Type|Account_frequency	0.7507643 0.4271785
6	Account_frequency|Transactions_Operation	0.9165218 0.8613892

6. Conclusion

The results of disparities of aggregate attributes score higher values for conventional function against the aggregate attribute with proposed function. We assume that disparities may affect on the model that may result and the Bayes Factor may indicate these effects, but the practical results reveals; there is no relationship between the disparities of aggregated values from the expected on the Bayes Factor. As a result we must be aware in using Bays Factor in evaluation models that based on aggregated attributes.

References

[1] Koller D. and Pfeffer A., "Probabilistic frame-based systems”, 1998, Proc. AAAI.

[2] Knobbe A. J., de Hass M., and Siebes A., "Propositionalisation and Aggregates", Principles of Data Mining and Knowledge Discovery, 5th Europea Conference, PKDD, Freiburg Germany, September 3-5 Preceding. Lectures Notes in Computer Science Springer, 2001, p.p.277-288. Koller D. and Pfeffer A., "Probabilistic frame-based systems”, 1998, Proc. AAAI.

[3] Dimitri B. P. and John T. N., "Introduction to Probability", 2002, Athena Scientific, Belmont, Massachusetts.

[4] Kass, R. and Raftery, “Bayes factors and model uncertainty”, Journal American Statistical Association, v.(90), 1995.

[5] Berka P., “Guide to the financial data set. The ECML/PKDD 2000 Discovery Challenge.

[6] Lamia, A., “Learning in Relational Data Mining”, A Dissertation Submitted to Iraqi Commission for Computers and Informatics/ Informatics Institute for Postgraduate Studies in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Computer Science, 2007.