Resolving Open Problems in Query Optimization Using Pattern Classification Techniques

B. John Oommen*[*] and Luis G. Rueda[†]

Abstract

We have solved the following problem using Pattern Classification Techniques (PCT): Given two histogram methods M_1 and M_2 used in *query optimization*, if the estimation accuracy of M_1 is greater than that of M_2, then M_1 has a higher probability of leading to the optimal Query Evaluation Plan (QEP) than M_2. To the best of our knowledge, this problem has been open for at least two decades, the difficulty of the problem partially being due to the hurdles involved in the formulation itself. By formulating the problem from a Pattern Recognition (PR) perspective, we use PCT to present a mathematical, rigorous proof of this fact, and show some uniqueness results. We also report empirical results demonstrating the power of these theoretical results on well-known histogram estimation methods.

Keywords: *Pattern Classification in Databases, Query Optimization, Histogram Methods.*

1 Introduction

1.1 Problem Statement

The theory of Pattern Recognition (PR) is quite advanced. Numerous books and papers have been written to present a foundational basis for the field [1, 2]. As opposed to this, the area of *query optimization* in database technology has still quite a few open, unsolved problems. In this short paper, we show that a fundamental open problem in the database theory can be solved using the principles of the theory of *pattern classification.*

Query optimization is an NP-Hard problem [3, 4] which has been studied for many decades. Given a query, the main problem is to find the optimal Query Evaluation Plan (QEP). In this paper we resolve a problem, which has been (to our knowledge) open[1] for more than two decades and describes how the accuracy of an estimation method relates to the quality of the solution obtained. More specifically, we solve the following

[*] Senior Member, IEEE. School of Computer Science, Carleton University, 1125 Colonel By Dr., Ottawa, ON, K1S 5B6, Canada. E-mail: oommen@scs.carleton.ca. Partially supported by NSERC, the Natural Science and Engineering Research Council of Canada.

[†] School of Computer Science, Carleton University, 1125 Colonel By Dr., Ottawa, ON, K1S 5B6, Canada. E-mail: lrueda@scs.carleton.ca. Partially supported by Departamento de Informática, Universidad Nacional de San Juan, Argentina.

[1] With reference to the open problem, please see the next subsection.

problem: Given two histogram methods M_1 and M_2 used in Query Optimization, if the estimation accuracy of M_1 is greater than that of M_2, then M_1 has a higher probability of leading to the optimal QEP than M_2. By means of a rigorous analysis based on a distribution used in the failure models, the *doubly exponential* distribution, we prove that the higher the accuracy of the method, the greater is the probability of this method yielding the optimal solution. The difficulty of this open problem is partially being due to the hurdles involved in the formulation itself. Indeed, after formulating the problem from a PR perspective, we use *pattern classification* techniques to present a mathematical, rigorous proof of this fact, and show some uniqueness results.

1.2 Importance of the Result

In the 1999 IDEAS Conference in Montreal, Canada, the database authority, Prof. J. D. Ullman from Stanford proposed the following question. He queried: "Does a system using a superior histogram method necessarily yield a superior QEP?". He also alluded to the experimental results using the Equi-width [5, 6] and Equi-depth histograms [7, 8] which *seemed to imply* that the answer to the query was *negative*.

The importance of the results of this paper is that we show that the answer to his question is "stochastically positive". In other words, we prove that although a superior histogram method may not always yield a superior QEP, the probability that the superior histogram method yields a superior QEP exceeds the problem that it yields an inferior QEP. This thus justifies and gives a formal rigorous basis for the fact that all current day database systems use histogram methods to determine the QEP.

We also show that if two "almost comparable" histogram methods (like the Equi-width and the Equi-depth) are compared, the probability of the superior one yielding a superior QEP is negligible. This probably answers for Prof. Ullman's implied position. However, if the error of one method is *significantly* less than the error of the second, the probability of obtaining a superior QEP is also significantly greater. This is because of the explicit form of the function involved, and further answers for the experimental results alluded to by Prof. Ullman. The corresponding result for significantly superior histograms also follows from the functional form, and is verified by experimental results involving the Equi-width, the Equi-depth, and the Rectangular Cardinality Attribute Cardinality Map (R-ACM), a recently devised histogram method [9].

The theoretical significance of this paper is the fact that it demonstrates how PR techniques can be used to solve open "unsolved" problems in various other unrelated domains. It also shows that certain decision problems do not necessarily have YES/NO answers, but answers which are stochastically positive/negative. In particular, we present a solution to this fundamental, unsolved problem in the area of query optimization in database technology using the theory of pattern classification.

1.3 Overview

We consider the fundamental query optimization problem. When an end user performs a query, many internal operations need to be done to retrieve the information requested. The most important operation between tables is the natural join on a particular attribute. In real databases, a query may consist of joining several tables. When more than two tables have to be joined, intermediate join operations are performed to ultimately obtain the final relation. As a result, the same query can be performed by means of different intermediate (join) operations. A simple sequence of join operations that leads to the same final result is called a QEP. Each QEP has associated an internal cost, which depends on the number of operations performed in the intermediate joins. The problem of choosing the best QEP is a combinatorially explosive optimization problem. This problem is currently solved by estimating the query result sizes of the intermediate relations and selecting the most efficient access QEP.

Since the analysis of selecting the best QEP must be done in "real" time, it is not possible to inspect the real data in this phase. Consequently, query result sizes are usually estimated using statistical information about the structures and the data maintained in the database catalogue. This information is used to approximate the distribution of the attribute values in a particular relation. Hence the problem of selecting the best QEP depends on how well that distribution is approximated.

In [10], it has been shown that errors in query result size estimates may increase exponentially with the number of joins. Since current databases and the associated queries increase in complexity, numerous efforts have being made to devise more efficient techniques that solve the query optimization problem.

Many techniques have been proposed to estimate query result sizes, including histograms, sampling, and parametric techniques [5, 7, 8, 11]. Histograms are the most commonly used form of statistical information. They are incorporated in most of the commercial database systems such as Oracle, Microsoft SQL Server, Teradata, and DB2, which mainly use the Equi-depth histogram. The prominent models of histograms known in the literature are: *Equi-width* [5, 6], *Equi-Depth* [7, 8], the *R-ACM* [9], the *Trapezoidal Attribute Cardinality Map (T-ACM)* [12], and the *V-Optimal Histograms* [10, 13].

In this paper, we focus on these histogram methods (or for that matter any histogram estimation methods). We analytically prove that under certain models, the better the accuracy of an estimation technique, the greater the probability of it choosing the optimal QEP.

In order to provide additional evidence, we have also provided some empirical results that shows the superiority of R-ACM over the traditional histogram estimation methods, the Equi-width and the Equi-depth. The empirical results obtained by testing these properties for many of the above histogram methods in random databases show that the R-ACM is significantly superior to both the Equi-width and the Equi-depth schemes.

2 The Relation between Efficiency and Optimality

Consider two query-size estimation methods, M_1 and M_2. The probability of choosing a cost value of a particular QEP by M_1 and that of choosing a cost value by M_2 are represented by two independent random variables. Clearly, this assumption of independence is valid because there is no reason why the value obtained by one estimation strategy should affect the value obtained by the second.

Although we consider the analysis for any two arbitrary histogram schemes, referred to by M_1 and M_2,

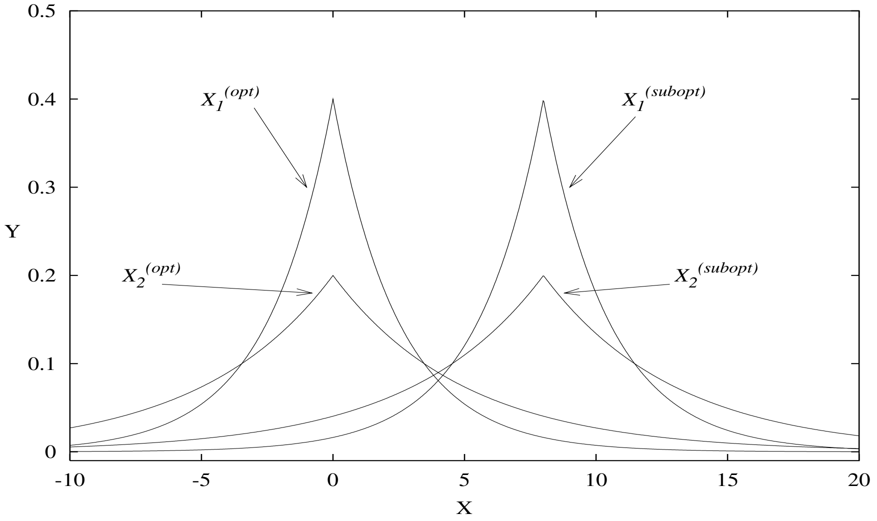

Figure 1: An example of doubly exponential distributions for the random variables $X_1^{(opt)}$, $X_2^{(opt)}$, $X_1^{(subopt)}$ and $X_2^{(subopt)}$, whose parameters are $\lambda_1 = 0.4$ and $\lambda_2 = 0.2$.

the result we claim can be extended to any of the query-size estimation methods mentioned above.

For the analysis done below, we work with the model that the error function is doubly exponential. In other words, the probability of obtaining a value that deviates from the mean falls exponentially as a function of the deviation. This model (unlike the Gaussian error function) is more typical in reliability analysis and in failure models, and in this particular domain, the question is one of evaluating how reliable the quality of a QEP is if only an estimate of its performance is available.

Without loss of generality, if the mean cost of the optimal QEP is μ, by shifting the origin by μ, we can work with the assumption that the cost of the best QEP is 0, which is the mean of these two random variables. The cost of the second best QEP is given by another two random variables (one for M_1 and the other one for M_2) whose mean, $c > 0$, is the same for both variables. An example will help to clarify this.

Example 1. Suppose that M_1 chooses the optimal cost value with probability represented by the random variable $X_1^{(opt)}$ whose mean is 0 and $\lambda_1 = 0.4$. This method also chooses another sub-optimal cost value according to $X_1^{(subopt)}$ whose mean is 8 and $\lambda_1 = 0.4$.

M_2 is another method that chooses the optimal cost value with probability given by $X_2^{(opt)}$ whose parameters are $M = 0$ and $\lambda_2 = 0.2$. Another sub-optimal cost value is chose with probability given by $X_2^{(subopt)}$ whose parameters are $M = 8$ and $\lambda_2 = 0.2$.

Since $\frac{2}{\lambda_1^2} < \frac{2}{\lambda_2^2}$, we would hope that the probability that M_1 chooses a sub-optimal cost value is smaller than that of M_2 choosing the sub-optimal cost value. This scenario is depicted in Figure 1. The result depicted above is formalized in the following theorem, *which is the primary result of this short paper, and answers the open question referred to above.* Observe too that the formulation and proof use techniques typically foreign to database theory, but which are fundamental to the theory of PR. □

Theorem 1. Suppose that:

- M_1 and M_2 are two query result size estimation methods.
- X_1 and X_2 are two doubly exponential random variables that represent the cost values of the *optimal* QEP obtained by M_1 and M_2 respectively.
- X_1' and X_2' are another two random variables representing the cost value of a *non-optimal* QEP obtained by M_1 and M_2 respectively.
- $0 = E[X_1] = E[X_2] \leq E[X_1'] = E[X_2'] = c$.

Let p_1 and p_2 be the probabilities that M_1 and M_2 respectively make the wrong decision. Then,

$$\text{if } \mathrm{Var}[X_1] = \mathrm{Var}[X_1'] = \frac{2}{\lambda_1^2} \leq \frac{2}{\lambda_2^2} = \mathrm{Var}[X_2] = \mathrm{Var}[X_2'], \quad p_1 \leq p_2 \ .$$

Proof. Consider a particular value x. The probability that the value x leads to a wrong decision made by M_1, is given by:

$$\begin{aligned} I_{11} &= \textstyle\int_{-\infty}^{x} \frac{1}{2}\lambda_1 e^{\lambda_1(u-c)}\, du && \text{if } x < c \ , \text{ and} \\ I_{12} &= \textstyle\int_{-\infty}^{c} \frac{1}{2}\lambda_1 e^{\lambda_1(u-c)}\, du + \int_{c}^{x} \frac{1}{2}\lambda_1 e^{-\lambda_1(u-c)}\, du && \text{if } x > c \ . \end{aligned} \tag{1}$$

Solving the integrals, (1) results in:

$$\begin{aligned} I_{11} &= \tfrac{1}{2}e^{\lambda_1(x-c)} - \lim_{u\to-\infty} \tfrac{1}{2}e^{-\lambda_1(u-c)} && = \tfrac{1}{2}e^{-\lambda_1(x-c)} \ , \text{ and} \\ I_{12} &= \lim_{u\to-\infty} \tfrac{1}{2}e^{-\lambda_1(-u+c)} + \tfrac{1}{2} - \tfrac{1}{2}e^{-\lambda_1(x-c)} + \tfrac{1}{2} && = 1 - \tfrac{1}{2}e^{-\lambda_1(x-c)} \ . \end{aligned} \tag{2}$$

The probability that M_1 makes the wrong decision for *all* the values of x is the following function of λ_1 and c:

$$p_1 = I(\lambda_1, c) = \int_{-\infty}^{0} I_{11}\frac{1}{2}\lambda_1 e^{\lambda_1 x}\, dx + \int_{0}^{c} I_{11}\frac{1}{2}\lambda_1 e^{-\lambda_1 x}\, dx + \int_{c}^{\infty} I_{12}\frac{1}{2}\lambda_1 e^{-\lambda_1 x}\, dx \ . \tag{3}$$

which, after applying the distributive law and substituting the values of I_{11} and I_{12}, can be written as:

$$\int_{-\infty}^{0} \frac{\lambda_1}{4} e^{2\lambda_1 x - \lambda_1 c}\, dx - \int_{0}^{c} \frac{\lambda_1}{4} e^{-\lambda_1 c}\, dx + \int_{c}^{\infty} \frac{\lambda_1}{2} e^{-\lambda_1 x} - \frac{\lambda_1}{4} e^{-2\lambda_1 x + \lambda_1 c}\, dx \ . \tag{4}$$

After solving the integrals, (4) is transformed into:

$$\frac{1}{8} e^{-\lambda_1 c} + \frac{1}{4}\lambda_1 c e^{-\lambda_1 c} + \frac{3}{8} e^{-\lambda_1 c} = \frac{1}{2} e^{-\lambda_1 c} + \frac{1}{4}\lambda_1 c e^{-\lambda_1 c} \ . \tag{5}$$

Similarly, we do the same analysis for p_2, which is a function of λ_2 and c:

$$p_2 = I(\lambda_2, c) = \frac{1}{2} e^{-\lambda_2 c} + \frac{1}{4}\lambda_2 c e^{-\lambda_2 c} \ . \tag{6}$$

We have to prove that:

$$p_1 = \frac{1}{2} e^{-\lambda_1 c} + \frac{1}{4}\lambda_1 c e^{-\lambda_1 c} \le \frac{1}{2} e^{-\lambda_2 c} + \frac{1}{4}\lambda_2 c e^{-\lambda_2 c} = p_2 \ . \tag{7}$$

Multiplying both sides by 2, and substituting $\lambda_1 c$ for α_1 and $\lambda_2 c$ for α_2, (7) can be written as follows:

$$e^{-\alpha_1} + \frac{1}{2}\alpha_1 e^{-\alpha_1} \le e^{-\alpha_2} + \frac{1}{2}\alpha_2 e^{-\alpha_2} \ . \tag{8}$$

Substituting α_2 for $k\alpha_1$, $\alpha_1 \ge 0$ and $0 < k \le 1$, (8) results in:

$$q_1 = e^{-\alpha_1} + \frac{1}{2}\alpha_1 e^{-\alpha_1} \le e^{-k\alpha_1} + \frac{1}{2}k\alpha_1 e^{-k\alpha_1} = q_2 \, . \tag{9}$$

We now prove that $q_1 - q_2 \le 0$. After applying natural logarithm to both sides of (9) and some algebraic manipulation, $q_1 - q_2 \le 0$ implies:

$$F(\alpha_1, k) = k\alpha_1 - \alpha_1 + \ln(1 + \frac{1}{2}\alpha_1) - \ln(1 + \frac{1}{2}k\alpha_1) \le 0 \, . \tag{10}$$

To prove that $F(\alpha_1, k) \le 0$, we use the fact that $\ln x \le x - 1$. Hence, we have:

$$
\begin{aligned}
F(\alpha_1, k) &= \alpha_1(k-1) + \ln\left(\frac{1+\frac{1}{2}\alpha_1}{1+\frac{1}{2}k\alpha_1}\right) && (11)\\
&\leq \alpha_1(k-1) + \frac{1+\frac{1}{2}\alpha_1}{1+\frac{1}{2}k\alpha_1} - 1 && (12)\\
&= \alpha_1(k-1) + \frac{\alpha_1 - k\alpha_1}{2+k\alpha_1} && (13)\\
&= \frac{k\alpha_1 + k^2\alpha_1^2 - \alpha_1 - k\alpha_1^2}{2+k\alpha_1} && (14)\\
&= \frac{\alpha_1(k-1)(k\alpha_1+1)}{2+k\alpha_1} \leq 0\,, && (15)
\end{aligned}
$$

because:

(i) $0 < k \leq 1$ and $\alpha_1 \geq 0 \;\Rightarrow\; \alpha_1(k-1) \leq 0$ and $k\alpha_1 + 1 > 0$. Hence $\alpha_1(k-1)(k\alpha_1+1) \leq 0$, and

(ii) $0 < k \leq 1$ and $\alpha_1 \geq 0 \;\Rightarrow\; 0 < k\alpha_1 \leq \alpha_1 \;\Rightarrow\; k\alpha_1 + 2 > 2 > 0$.

Hence the theorem. □

The above theorem can be viewed as a "sufficiency result". In other words, we have shown that $q_1 - q_2 \leq 0$ or that $p_1 \leq p_2$. We now show a "necessity result" stated as a uniqueness result. This result states that the function $p_1 \leq p_2$ has its equality ONLY at the boundary condition where the two distributions are exactly identical.

To prove the necessity result, we consider $q_2 - q_1$ which, derived from (9), can be written, as a function of α_1 and k, as:

$$G(\alpha_1, k) = e^{-k\alpha_1} + \frac{1}{2}k\alpha_1 e^{-k\alpha_1} - e^{-\alpha_1} - \frac{1}{2}\alpha_1 e^{-\alpha_1}\,. \qquad (16)$$

By examining its partial derivatives we shall show that there are two solutions for equality. Furthermore, when $\alpha_1 \geq 0$ and $0 < k \leq 1$, we shall see that for a given k, there is only one solution, namely $\alpha_1 = 0$ and k, $0 < k \leq 1$, proving the uniqueness.

Theorem 2. Suppose that $\alpha_1 \geq 0$, $0 < k \leq 1$. Let $G(\alpha_1, k)$ be:

$$G(\alpha_1, k) = e^{-k\alpha_1} + \frac{1}{2}k\alpha_1 e^{-k\alpha_1} - e^{-\alpha_1} - \frac{1}{2}\alpha_1 e^{-\alpha_1}\,. \qquad (17)$$

Then $G(\alpha_1, k) \geq 0$, and there are exactly two solutions for $G(\alpha_1, k) = 0$, being: $\{\alpha_1 = -1, k = 1\}$ and $\{\alpha_1 = 0, k\}$.

Proof. We must prove that, as defined in the theorem statement, $G(\alpha_1, k) \geq 0$.

We shall prove that this is satisfied by determining the local minima for $G(.,.)$, where $\alpha_1 \geq 0$ and $0 < k \leq 1$. We first find the partial derivatives of (17) with respect to α_1 and k:

$$\frac{\partial G}{\partial \alpha_1} = -\tfrac{1}{2}ke^{-k\alpha_1} - \tfrac{1}{2}k^2\alpha_1 e^{-k\alpha_1} + \tfrac{1}{2}e^{-\alpha_1} + \tfrac{1}{2}\alpha_1 e^{-\alpha_1} = 0\,, \text{ and} \tag{18}$$

$$\frac{\partial G}{\partial k} = -\tfrac{1}{2}\alpha_1 e^{-k\alpha_1} - \tfrac{1}{2}k\alpha_1^2 e^{-k\alpha_1} = 0\,. \tag{19}$$

We now solve (18) and (19) for α_1 and k. Equation (19) can be written as follows:

$$-\frac{1}{2}\alpha_1 e^{-k\alpha_1} = \frac{1}{2}k\alpha_1^2 e^{-k\alpha_1}\,, \tag{20}$$

which, after canceling some terms results in $k\alpha_1^2 + \alpha_1 = 0$. Solving this equation for α_1, we have: $\alpha_1 = -\frac{1}{k}$ and $\alpha_1 = 0$. Substituting $\alpha_1 = -\frac{1}{k}$ in (18) and canceling some terms, we obtain:

$$\frac{1}{2}e^{-\alpha_1} + \frac{1}{2}\alpha_1 e^{-\alpha_1} = 0\,, \tag{21}$$

which results in the solution to be $\alpha_1 = -1$, and consequently, $k = 1$.

The second root, $\alpha_1 = 0$, indicates that the minimum is achieved for any value of k.

We have thus found two solutions for (18) and (19), $\{\alpha_1 = 0, k\}$ and $\{\alpha_1 = -1, k = 1\}$. Since $\alpha_1 \geq 0$, it means that α_1 can have at least a value of 0, and hence the local minima is in $\{\alpha_1 = 0, k\}$. Substituting these two values in G, we see that $G(\alpha_1, k) = 0$, which is the minimum. Therefore, $G(\alpha_1, k) \geq 0$ for $\alpha_1 \geq 0$ and $0 < k \leq 1$.

Hence the theorem. □

To get a physical perspective of these results, let us analyze the geometric relation of the function G and the histograms estimation methods. G is a positive function in the region $\alpha_1 \geq 0$, $0 < k \leq 1$. When $\alpha_1 \to 0$, $G \to 0$. This means that for small values of α_1, G is also small. Since $\alpha_1 = \lambda_1 c$, the value of α_1 depends on λ_1 and c. When c is small, G is very close to its minimum, 0, and hence both probabilities, p_1 and p_2, are very close. This behavior can be observed in Figure 2.

In terms of histogram methods and QEPs, when c is small, the optimal and the sub-optimal QEP are very close. Since histogram methods such as Equi-width and Equi-depth produce a larger error than the R-ACM and the T-ACM, the former are less likely to find the optimal QEP than the latter.

On the other hand, G is very small when λ_1 is close to 0. This means that $\mathrm{Var}[X_1]$ is very large. Since $\mathrm{Var}[X_1] \leq \mathrm{Var}[X_2]$, $\mathrm{Var}[X_2]$ is also very large, and both are close each other (In Figure 1, we would observe almost flat curves for both distributions). Random variables for histogram methods such as Equi-width and Equi-depth yield similar error estimation distributions with large and similar variances. Hence, the

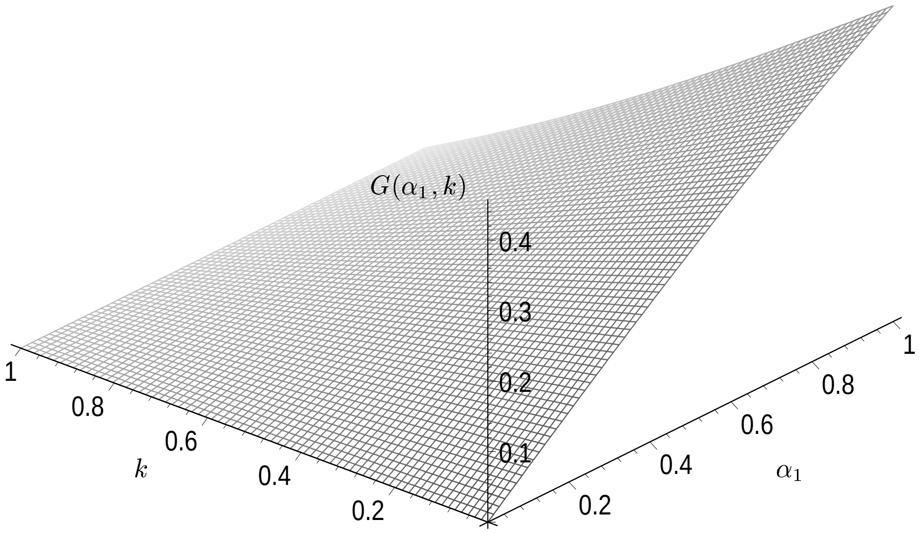

Figure 2: Function $G(\alpha_1, k)$ plotted in the ranges $0 \leq \alpha_1 \leq 1$ and $0 \leq k \leq 1$.

probabilities p_1 and p_2 are quite close, and consequently, similar results are expected for these estimation methods. As a consequence of this, the results of Theorems 1 and 2 are not too meaningful in the absence of the new histogram methods, the R-ACM and the T-ACM, which effectively imply random variables with smaller variances and with underlying random variables quite different from those implied by the Equi-width and the Equi-depth methods.

3 Empirical Results

In order to provide practical evidence of the theoretical results presented above[2], we have performed some simulations with randomly generated databases. In the experiments we have conducted, the details of which can be found in [14], four independent runs. In each run, 100 random databases were generated. Each database was composed of six relations, each of them having six attributes. Each relation was populated with 100 tuples.

The efficiency of R-ACM was compared with that of the Equi-width and the Equi-depth after performing these simulations using 50 values per attribute. We set the number of bins for the Equi-width and the Equi-depth to be 22. In order to be impartial with the evaluation, we set the number of bins for the R-ACM to be *approximately half* of that of the Equi-width and the Equi-depth, because the former needs twice as much storage as that of the latter.

The simulation results obtained from 400 independent runs, used to compare the efficiency of the R-ACM with that of the Equi-width and that of the Equi-depth, are given in Table 1. The column labeled "R-ACM-W"

[2] This paper is not intended to compare the various histogram methods: Equi-width, Equi-depth, R-ACM, T-ACM, V-optimal, etc. The experimental results submitted are merely included to demonstrate that the theoretically proven results can be experimentally justified.

Simulation	**R-ACM-W**	**Equi-width**	**R-ACM-D**	**Equi-depth**
1	26	12	35	12
2	24	15	42	13
3	35	11	46	8
4	29	15	46	8
Total	**114**	**53**	**169**	**41**

Table 1: Simulation results for the R-ACM, Equi-width, and Equi-depth, after optimizing a query on 400 randomly generated databases.

is the number of times that R-ACM is better than Equi-width. The column labelled "Equi-width" indicates the number of times in which the Equi-width obtains a better QEP than that of the R-ACM. Similarly, the column labelled "R-ACM-D" represents the number of times that the R-ACM yields better solutions than Equi-depth, and the column labelled "Equi-depth" is the number of times in which the Equi-depth is superior to the R-ACM. The last row, the total of each column gives us the evidence that the R-ACM is superior to Equi-width in more than twice as much, and the R-ACM is better than Equi-depth by a factor of about four.

4 Conclusions

The theory of PR is quite developed, and has many applications. We believe that this theory can be used to prove unsolved results in various other fields. In particular, we have applied pattern classification techniques to solve problems in the area of database query optimization. In this paper, we have discussed the efficiency of using histogram estimation methods for database query optimization and resolved an open problem, which has been (to our knowledge) open for at least twenty years. The problem describes how the accuracy of an estimation method relates to the quality of the solution obtained. The efficiency has been quantified by means of the probability of a method choosing the optimal solution.

We have shown analytically (using a reasonable model of accuracy, namely the doubly exponential distribution for errors) that as the accuracy of an estimation method increases, the probability of it leading to a superior QEP also increases. We have shown that histogram methods that produce errors with similar variances (such as the recently introduced R-ACM and T-ACM), the expected results are also similar. We have also shown that the R-ACM and the T-ACM, which produce error with smaller variances than the traditional methods, yield better QEPs in a substantially larger number of times.

We have also provided evidence of the theoretical results by means of the empirical results obtained from evaluating the Equi-width, the Equi-depth, and the R-ACM on randomly generated databases. These results show that the R-ACM provides superior solutions in more than twice as many times as the Equi-width, and in more than four times often than the Equi-depth.

The question of analyzing the accuracy/speed problem for other distributions (for example, Gaussian) remains open, but is far from trivial. More detailed empirical results including the design of random databases and random queries in these random databases can be found in [14].

References

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

[2] A. Webb, *Statistical Pattern Recognition.* New York: Oxford University Press Inc., 1999.

[3] S. Cluet and G. Moerkotte, "On the complexity of generating optimal left-deep processing trees with cartesian products," in *Proceedings of the International Conference on Databases Theory*, (Prague), pp. 54–67, 1995.

[4] W. Scheufele and G. Moerkotte, "On the complexity of generating optimal plans with cross products," in *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, (Tucson, Arizona), pp. 238–248, 1997.

[5] R. P. Kooi, *The optimization of queries in relational databases.* PhD thesis, Case Western Reserve University, 1980.

[6] S. Christodoulakis, "Estimating selectivities in data bases," in *Technical Report CSRG-136*, (Computer Science Dept, University of Toronto), 1981.

[7] G. Piatetsky-Shapiro and C. Connell, "Accurate estimation of the number of tuples satisfying a condition," in *Proceedings of ACM-SIGMOD Conference*, pp. 256–276, 1984.

[8] M. Muralikrishna and D. J. Dewitt, "Equi-depth histograms for estimating selectivity factors for multi-dimensional queries," in *Proceedings of ACM-SIGMOD Conference*, pp. 28–36, 1988.

[9] B. J. Oommen and M. Thiyagarajah, "The Rectangular Attribute Cardinality Map: A New Histogram-like Technique for Query Optimization," in *International Database Engineering and Applications Symposium, IDEAS'99*, (Montreal, Canada), pp. 3–15, August 1999.

[10] Y. Ioannidis and S. Christodoulakis, "On the propagation of errors in the size of join results," in *Proceedings of the ACM-SIGMOD Conference*, pp. 268–277, 1991.

[11] M. Mannino, P. Chu, and T. Sager, "Statistical profile estimation in database systems," in *ACM Computing Surveys*, vol. 20, pp. 192–221, 1988.

[12] B. J. Oommen and M. Thiyagarajah, "The Trapezoidal Attribute Cardinality Map: A New Histogram-like Technique for Query Optimization," Tech. Rep. TR-99-04, School of Computer Science, Carleton University, Ottawa, Canada, February 1999.

[13] W. Poosala, *Histogram Based Estimation Techniques in Databases.* PhD thesis, University of Wisconsin - Madison, 1997.

[14] B. J. Oommen and L. G. Rueda, "The Efficiency of Modern-day Histogram-like Techniques for Query Optimization," *Submitted for Publication.*