Efficient probability density balancing for supporting distributed knowledge discovery in large databases*

Dragan Obradovic[1] and Zoran Obradovic[2]
dragan.obradovic@mchp.siemens.de and zoran@eecs.wsu.edu

[1]Siemens AG, Corporate Technology, Information and Communications,
Otto-Hahn-Ring 6, 81739 Munich, Germany
[2]School of Electrical Engineering and Computer Science,
Washington State University, Pullman, WA 99164, USA

Abstract

Most of the existing nonlinear data analysis and modelling techniques including neural networks become computationally prohibitively expensive when the available data set exceeds the capacity of the computer main memory due to the slow disc access operations [1]. For the data received on-line from a source with an unknown probability distribution, the question addressed in this article is how to efficiently partition it to smaller representative subsets (data bases) and how to organize these data subsets in order to minimize the computational cost of the later data analysis. The proposed linear-time, on-line problem decomposition method achieves these objectives through balancing probability distributions of the individual disjoint data subsets, each aimed at approximating the original data-source distribution. Consequently, computationally efficient statistical data analysis and neural network modelling on data subsets fitting into a computer central memory will produce results similar to these obtained through a global, computationally infeasible data analysis. In addition, the proposed decomposition scheme enables for an effective distributed data analysis on a network of workstations (a fixed or an adaptive size) since different modeling algorithms can be run simultaneously on disjoint data subsets with no data exchange and with minimal communication of higher-level locally obtained knowledge.

Purpose

Data mining and decision making based on the distributed data-bases are becoming increasingly important for both scientific and economic reasons [2]. The availability of huge data sets and the demand for a decentralized decision making have resulted in the need for the optimization of the distributed data bases [3].

In this paper we assume that the data is generated by an unknown distribution and then send in blocks to different data bases. Moreover, we assume that the data selection is biased. This means that the bock of the N data to be sent is not chosen randomly from the original data-pool but was subject to deterministic ordering. A simple theoretical example is a case where data corresponding to a Gaussian distribution are ordered and then the first third of the data sent to the first data base, second third to the second data base and remaining third to the last data base. This data selection will result in data distributions that differ from each other and from the original Gaussian distribution. Consequently, any statistical data mining or data decision process based on any individual data-base will be biased. The bias will be avoided only if access to all three data bases is possible, which might be infeasible or computationally too expensive. Another simple but practical situation with potential biased sampling problems is a customer database of a major medical insurance company collecting claim data from pediatrics, family practice, adult health care and other clinics, where clients age distribution is

*Partial support by the INEEL University Research Consortium grant No.C94-175936 and the NSF research grant NSF-CSE-IIS-9711532 to Z. Obradovic is gratefully acknowledged.

0-7803-5529-6/99/$10.00 ©1999 IEEE

clearly different for data stemming from various sites (data bases). In order to address such problems, we derive a novel data-assignment mechanism based on the information theoretic postulates that minimizes the difference between the individual data distribution functions.

Method

Let there be M data bases and let the block of data to be sent be of size N. Furthermore, let the data range be known which will enable construction of bins for the histogram evaluation. If a probability density function of the i-th data base is $p_i(x)$, then the difference between the two distributions $p_i(x)$ and $p_j(x)$ is defined by the Kullback-Leibler distance [4]:

$$K(p_i, p_j) = \int p_i(x) \log \frac{p_i(x)}{p_j(x)} dx$$

The Kullback-Leibler distance is a semi measure, i.e. it is equal to zero if the two distributions are equal, it is always positive if distributions differ but it does not satisfy the triangular inequality. Moreover, the Kullback-Leibler measure is non-symmetric, i.e. $K(p_i,p_j) \neq K(p_j,p_i)$.

In order to avoid the problem of non-symmetry, we will construct a new distance K_1:

$$K_1(p_i, p_j) = K(p_i, p_j) + K(p_j, p_i) =$$

$$= \int p_i(x) \log \frac{p_i(x)}{p_j(x)} dx + \int p_j(x) \log \frac{p_j(x)}{p_i(x)} dx$$

The resulting distance is also non-negative but it is symmetric. The remaining thing to be done is to construct a variable that evaluates the differences among all M data distributions. A valid candidate is a sum of all the pairwise distances:

$$J(p_1, \ldots, p_M) = \sum_{i=1}^{M} \sum_{j \neq i}^{M} K_1(p_i, p_j)$$

Since the actual probability density function is unknown, the Kullback-Leibler distance has to be estimated from the data. A straightforward estimation is obtained by approximating the continuous density function $p_i(x)$ with a histogram having the prespecified bin structure corresponding to the known range of data. Hence, the estimate of the overall cost function $J(p_1,\ldots,p_M)$ will depend on the quality of the available histograms. In addition, potential problems might arise when there is a bin in one of the distributions p_j without any data since in such a case, the corresponding value of the logarithm will become infinite. In order to avoid this over-sensitivity of the measure, we initially assigned a single data point to every bin.

A flow-chart of the algorithm for a prespecified number of data bases is depicted in Figure 1 where data points stored to a single data base are indexed based on the bins they belong to. The algorithm is easily extendable to a scenario with an adaptive number of databases growing based on on-line needs (volume of the input data stream and physical limits of desired data subsets. In such a case, an index to each bin is implemented as a double hash data structure with the first hash key for a bin pointing to the begining and the second to the middle of a corresponding data list. When a single data base overgrows the prespecified capacity (e.g. the computer central memory size, or an external storage device size) this index structure is used to efficiently split the oversized set into two equal-size, similar distribution subsets by moving halph of the data from each bin to a new subset on the same external storage device.

Results

An important characteristic of the proposed method is that the variable x can be multidimensional as well as conditioned on some other variable y. Consequently, our algorithm is capable of optimally assigning multidimensional and/or conditional data to several data bases.

To store a new data block to an appropriate data base, it requires computing order of M^2 K_1 distances each obtained in a constant-time, N inserts to a single data base and possibly extra time to split an oversized data base which can be done in time linear in the size of such a data base. The number of available databases M as well as a block size is typically a constant, and so the cost of a single data block assignment for a fixed M is a small constant in practice.

The proposed data assignment algorithm is tested on a Gaussian distribution containing 2000 data points. It is assumed that the number of data bases is M=3 while the length of the data block was chosen to be N=5. The center of the bins are fixed at positions:[-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2].

The initial assumption was that the data is ordered in ascending order and the blocks are sequentially taken from the smallest value on. First, we considered the

naive assignment where the blocks are sent initially to the first, then to the second and at the end to the third data base so that each of them contains one third of the data. The corresponding histograms are depicted in Figure 2. It is easy to see that the corresponding distributions are very much different from the original Gaussian distribution and that, therefore, any statistical analysis of the individual data subsets would lead to the wrong understanding of the original data set. The properties of the original data set can be reliably estimated only by combining the data from all three data bases which is very expensive (all the data have to be copied from one place to another).

On the other hand, the results of the herein proposed algorithm are depicted in Figure 3. It is evident that the distributions of all three constructed data bases are approximately Gaussian. Hence, it suffices to analyze a single data base to recover the properties of the original data set in spite of the fact that its distribution was assumed to be unknown (the algorithm does not require knowledge of the original distribution, it only minimizes the differences between distributions corresponding to individual data bases).

New aspects of work

The introduced algorithm automatically and efficiently assigns the incoming data blocks to multiple databases so that the the corresponding data distributions are kept as close as possible to each other according to the modified Kullback-Leibler distance. The knowledge of the original distribution is not required.

Conclusions

The proposed on-line problem decomposition method achieves a near optimal performance through balancing probability distributions of the individual disjoint data subsets, each aimed at approximating the original data-source distribution. Consequently, computationally efficient statistical data analysis and neural network modelling on data subsets fitting into a computer central memory produce results similar to these obtained through a global, computationally infeasible data analysis. In addition, the proposed decomposition scheme enables an effective distributed data analysis on a network of workstations, since the different modeling algorithms can be run simultaneously on disjoint data sets with no data exchange and with minimal communication of higher-level locally obtained knowledge.

References

[1] Provost, F. and Kolluri, V, *A survey of methods for scaling up inductive algorithms,* in press.

[2] Chan, P. and Stolfo, S. "On the accuracy of meta-learning for scalable data mining," *J. Intelligent Information Systems,* vol. 8, 1997, pp. 5-28.

[3] Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P., "From data mining to knowledge discovery: An overview," *Advances in Knowledge Discovery and Data Mining,* AAAI Press, Menlo Park, CA, 1996.

[4] Cover, T.M. and Thomas, J.A.: *Elements of information theory*, John Wiley & Sons, New York, 1991.

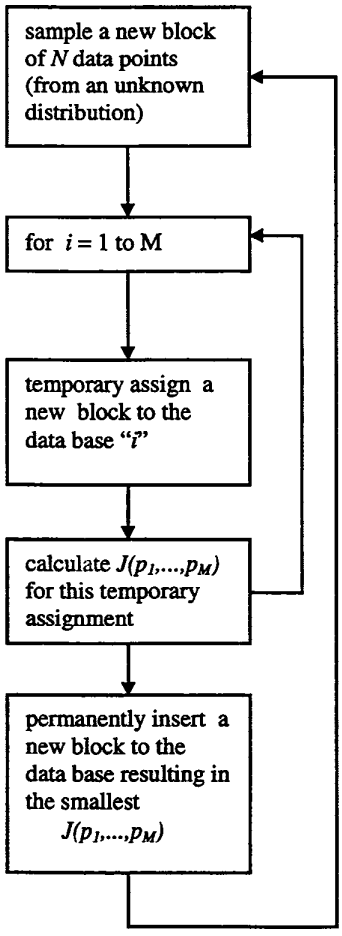

Figure 1. Flow chart of the assignment algorithm for a prespecified number of data bases

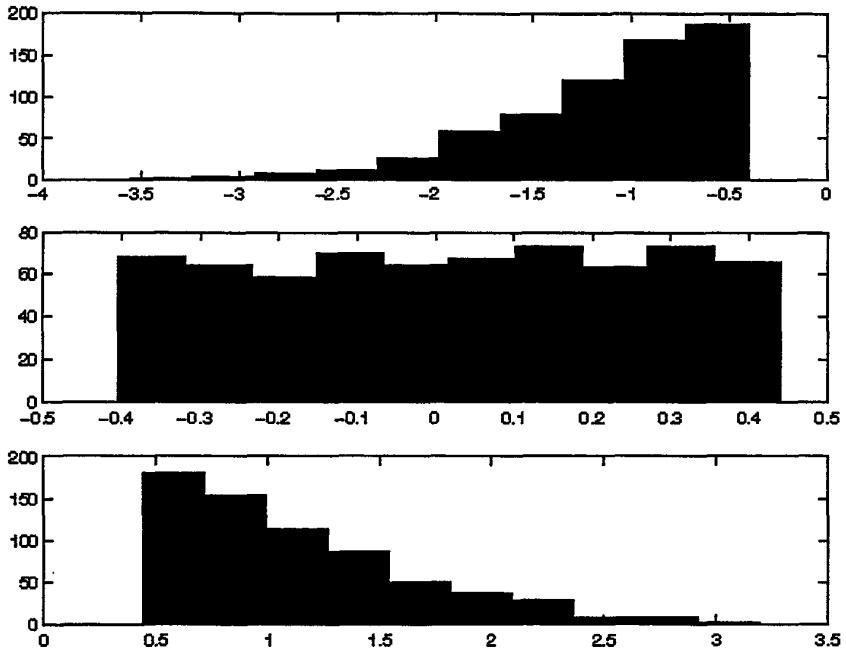

Figure 2. Histograms of the 3 data bases after the "naive" assignment

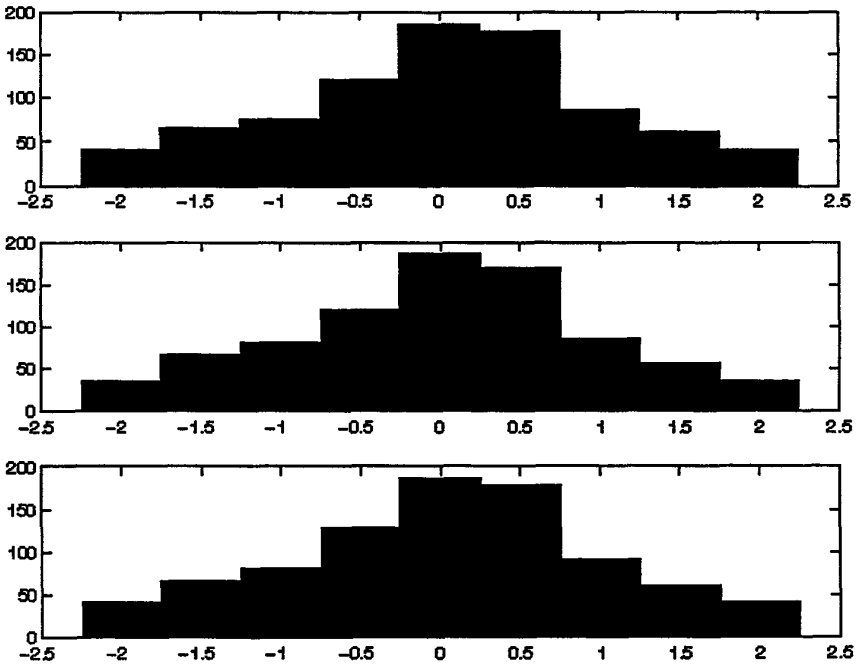

Figure 3. Histograms of the resulting data bases after applying the assignment algorithm