New Optimization Algorithm for Data Clustering

Jumi Kim
Korea Small Business Institute
(jmkim@kosbi.re.kr)

Large data handling is one of critical issues that the data mining community faces. This is particularly true for computationally intense tasks such as data clustering. Random sampling of instances is one possible means of achieving large data handling, but a pervasive problem with this approach is how to deal with the noise in the evaluation of the learning algorithm. This paper develops a new optimization based clustering approach using an algorithm specifically designed for noisy performance. Numerical results show this algorithm better than the other algorithms such as PAM and CLARA. Also with this algorithm substantial benefits can be achieved in terms of computational time without sacrificing solution quality using partial data.

Key Words : Data Clustering, Optimization-Based Clustering, NP Method, Partitioning, Random Sampling,

Received : May 2007 Accepted : August 2007 Corresponding author : Jumi Kim

1. Introduction

In recent years databases in modern enterprises have become massive and contain a wealth of important data. However when traditional methods of analysis fall short in transforming this data into knowledge, exploratory techniques such as knowledge discovery in databases must be applied. This multidisciplinary field of data mining draws heavily on statistics and artificial intelligence, but numerous problems in data mining and knowledge discovery can also be formulated as optimization problems, and optimization techniques can therefore be used to solve large-scale data mining problems (Basu, 1998; Bradley et al., 1999).

As the importance of data mining has grown, one of the critical issues to emerge is how to scale data mining techniques to larger databases (Bradley et al., 2002). This is particularly true for computationally intensive data mining tasks such as identifying natural clusters of instances (Kaufman and Rousseeuw, 1990). Several approaches to scalability enhancements have been studied at length in the literature (Provost and Kolluri, 1999), including using parallel mining algorithms (Forman and Zhang, 2000) and preprocessing the data by filtering out redundant or irrelevant features and thus reducing the dimensionality of the database (Ólafsson, 2003). Another approach to better scalability is using a selection of instance from a database rather than the entire database (Liu and Motoda, 2001). This paper deals with such instance selection and how it can be applied to data clustering within an optimization-based framework.

Perhaps the simplest approach to instance selection is random sampling (Kiven and Mannila, 1994). Numerous authors have studied this approach for specific data mining tasks such as clustering (Kaufman and Rousseeuw, 1990 ; Ng and Han, 1994, Ester, elt al., 1995), association rule discovery (Toivonen, 1996), and decision tree induction (Chauchat and Rakotomalala, 2001). When this approach is implemented the most challenging issue is determining a sample size that improves the performance of the algorithm without sacrificing the solution quality. Bounds can be developed that allow for a prediction of sample effort needed, but such bounds usually require knowing certain problem parameters and typically overestimate the necessary sample size (Toivonen, 1996). On the other hand, too small sample will lead to a bias and degeneration in performance. One possible solution is to use adaptive sampling (Domingo et al., 2002 ; Provost et al., 1999).

In this paper we advocate an alternative approach that is based on a novel formulation of the clustering task as an optimization problem. We also take advantage of the fact that certain optimization techniques have been explicitly designed to account for noisy performance estimates, which are common when performance is estimated using simulation. In particular, one such method is the nested partitions method that can be used to solve general global optimization problems (Shi and Olafsson, 2000) and specifically combinatorial type optimization problems with noisy performance (Olafsson, 1999). A characteristic of this method is that wrong moves made due to noise in performance estimates can be automatically corrected in a later move. In the scalable clustering context this means that noisy performance estimates, resulting from smaller samples of instance, may result in more steps taken by the algorithm but any bias will be automatically corrected. This eliminates the need to determine the exact sample size, although the computational performance of the algorithm may still depend on some extent on how it is selected.

The remainder of this paper is organized as follows. In Section 2 we briefly review clustering techniques and in particular, focus on efforts in scalable clustering. In Section 3 we discuss the basis for the new clustering methodology, which is an optimization method called the Nested Partitions method, present the optimization-based clustering algorithm which is called NPCLUSTER algorithm and demonstrate its effectiveness on a sample problem. In Section 4 we present some numerical results of the scalability of the algorithm with respect to the instance dimension, and Section 5 contains concluding remarks and suggestions for future research directions.

2. Scalable Clustering

Clustering has been an active area of research for several decades, and many clustering algorithms have been proposed in the literature (Kaufman and Rousseeuw, 1990; Grabmeier and Rudolph, 2002; Yim and Oh, 2003). In particular, considerable research has been devoted specifically to scalable clustering. We will start by briefly describing the various types of clustering algorithms and then mention some specific scalable methods.

Clustering algorithms can be roughly divided into two categories: hierarchical clustering and partitional clustering (Jain, et al., 1999). In hierarchical clustering all of the instances are organized into a hierarchy that describes the degree of similarity between those instances (e.g., a dendrogram). Such representation may provide a great deal of information, but the scalability of this approach is questionable as the number of instances grows. Partitional clustering, on the other hand, simply creates one partition of the data where each instance falls into one cluster. Thus, less information is obtained but the ability to deal with a large number of instances is improved. Examples of the partitioning approach are the classic k-means and k-medoids clustering algorithms.

There are many other characteristics of clustering algorithms that must be considered to ensure scalability of the approach. For example, most clustering algorithms are polythetic, meaning that all features are considered simultaneously in tasks so as to determine the similarity of two instances. However, as the number of features becomes high this may pose scalability problems and it may be necessary to restrict attention to monothetic clustering algorithms that consider one feature at a time. Most clustering algorithms are also non-incremental in the sense that all of the instances are considered simultaneously.

However, there are a few algorithms that are incremental, which implies that they consider each instance separately. Such algorithms are particularly useful when the number of instances is large.

Scalable clustering has received considerable attention in recent years, and here we will mention only a few of the methods that have been developed. For example, Zhang et al., (1996) proposed BIRCH, a hierarchical algorithm for clustering. The key idea of this method is to summarize cluster representations using two innovative concepts, clustering feature and clustering feature tree.

Another approach to hierarchical clustering is the CURE algorithm developed by Guha et al., (1998). The steps of the CURE algorithm followings; obtain a sample from the original database, partition the sample into a set of partitions and then cluster each partition, eliminate outliers and cluster the partial clusters. Finally, each data instance is labeled with the corresponding cluster.

Improved scalable versions of partitioning methods such as k-means and k-medoids have also been developed. The Clustering LARge Applications (CLARA) algorithm improves the scalability of the PAM k-medoids algorithm by applying PAM to multiple samples of the actual data and returns the best clustering (Kaufman and Rousseeuw, 1990).

A single pass k-means clustering algorithm was proposed by Bradley et al., (1998), with the main idea to use a buffer to save points from the database in a compressed form. This approach was simplified in the algorithm proposed by Farnstrom et al., (2000), in an effort to reduce the overhead that otherwise might cancel out any scalability improvements that might be achieved.

Yet another way of improving scalability is via distributed clustering, where instead of combining all data before clustering, data sets are operated on independently with minimum communication between the parallel clustering algorithms (Forman and Zhang, 2000).

The work presented in this paper is a partitional clustering algorithm that attempts to find cluster centers and uses random sampling to improve scalability. In that sense, it is the most similar to the CLARA algorithm, but its optimization-based approach sets it apart.

3. Optimization-Based Clustering

3.1 The NP-Method

The nested partitions (NP) method is an optimization method that has been suggested by Shi and Olafsson (2000) to solve general global optimization problems of the following form:

$$\min_{x \in X} f(x) \tag{1}$$

where x is a point in a n-dimensional space X and $f: X \rightarrow \mathbf{R}$ is a real-valued performance measure defined on this space. This performance may or may not be known deterministic. In our context, X is the space of all clusters and measures some quality of the clusters.

The intuitive idea of the NP method is quite simple. In each step, the method systematically partitions the feasible region into subsets and focuses the computational effort in those subsets that are considered promising. The main components of the method are:

- **Partitioning** : at each iteration the feasible region is partitioned into subsets that cover the feasible region but concentrate the search in what is believed to be the most promising region.
- **Random sampling** : to evaluate each of the subsets, a random sample of solutions are obtained from each subset and used to estimate the performance of the region as a whole.

This method can be understood as an optimization framework that combines adaptive global sampling with local heuristic search. It uses a flexible partitioning method to divide the design space into regions that can be analyzed individually and then aggregates the results from each region to determine how to continue the search, that is, to concentrate on the computational effort. Thus, the NP method adaptively samples from the entire design space and concentrates on the sampling effort by systematic partitioning of the design space.

To implement the partitioning, the NP method maintains in the kth iteration what is called the most promising region, that is, a subregion $X(k) \subseteq X$ that is considered the most likely to contain the best solution. This most promising region is partitioned into a given number of subregions and what remains is aggregated into one region called the surrounding region. Thus, a disjoint collection of sets covering the entire feasible region is considered. The subregions and the surrounding region are sampled using random sampling, and the sampling information used to determine which region should be the most promising region in the next iteration. If one of the subregions contains the best solution, this region is now selected as the new most promising region and is, in the next iteration, partitioned into smaller subregions. If the surrounding region contains the best solution this is taken as an indication that the last move might not have been the best move, so the algorithm backtracks to what was the most promising region in the previous iteration. This partitioning creates a tree of subsets that we refer to as the partitioning tree.

3.2 Defining Clusters

The partitional clustering problem can be formulated as an optimization problem and thus solved within the NP framework. In particular, we have designed the NP method as a partitional clustering method for nominal data and incorporated *k*-means into the same framework. In this approach we assume that we want to partition a given data set into *m* clusters and that each clusters is defined by its center (each instance is assigned to the closest center). The decision variables are thus the i-th coordinate of the *j*-th cluster, where $i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, m$. Therefore, this clustering problem reduces to locating the centers to optimize certain performance.

Selecting a performance measure to be optimized is very subjective, since determining what constitutes a good cluster is necessarily subjective and no single standard exists. We refer the reader to Estivill-Castro (2002) for a recent discussion of this issue and Grabmeyer and Rudolph (2002) for a more extensive survey. The most common measures are probably to maximize similarity within a cluster (that is, maximize homogeneity or compactness), and to minimize similarity between different clusters (that is, maximize separability between the clusters). A particular strength of the optimization-based framework is that any such measure, or combination of measures, can be adopted. Indeed, the function *f* can be defined as any measure of what is believed to indicate the quality of a cluster.

To minimize bias that may be introduced by analyzing very specific performance measures, we restrict ourselves here to a single measure of similarity within cluster, namely, its compactness :

$$f(\mathrm{x}) = \sum_{j=1}^{\mathrm{m}} \sum_{i=1}^{\mathrm{n}} \left| y_i^j - x_j \right|^2 \tag{2}$$

Here Θ is the space of all instances, $y_i^j \in \Theta$ is a specific instance in the space, x is the cluster center to which the instance is assigned, where $|y_i^j - x_j|^2$is the difference between a data point the y_i^j and cluster center x_j. So the objective function is an indicator of the distance of n data points from their respective cluster centers.

We believe that by using such a simple measure we are better able to focus on the performance of the algorithm itself. For a particular application, however, this will without doubt be defined in a different fashion, but that will not change the implementation of the algorithm.

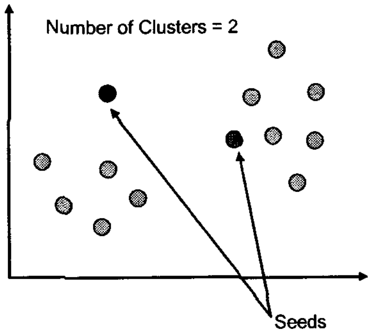
3.3 Partitioning

The main implementation issue for applying NP is to define the partitioning. We suggest doing this by finding cluster centers for one feature at a time, that is, at each level of the partitioning tree the values for all centers are limited to a given range for one feature. This defines the subsets or regions that form the partitioning tree. Then, as for the generic NP method, random samples are obtained from each subset, and to speed convergence, the k-means algorithm is applied to those random samples and the resulting improved centers used to select the most promising region. This most promising region is partitioned further, the surrounding region is aggregated, and so forth.

To help clear understanding, let's see the simple example. To present a detail description, implementation of the NPCLUSTER method we need the following notation:

Θ	: The space of all instances
$\sigma(k)$	: The most promising region in the k-th iteration
Σ	: $\{\sigma \subseteq \Theta \mid \sigma$ is a valid region given a fixed partitioning $\}$
Σ_0	: $\{\sigma \subseteq \Sigma \mid \sigma$ is a maximum depth $\}$
$s(\sigma)$	: The super-region of $\sigma \in \Sigma$
$d(\sigma)$	: Depth of region σ
d^*	: Maximum depth
m	: Total number of clusters(given)
n	: Total number of features(given)

[Figure 1] shows simple example for clustering using NP methodology. This is nominal problem with two dimensions. The total number of cluster is 2.



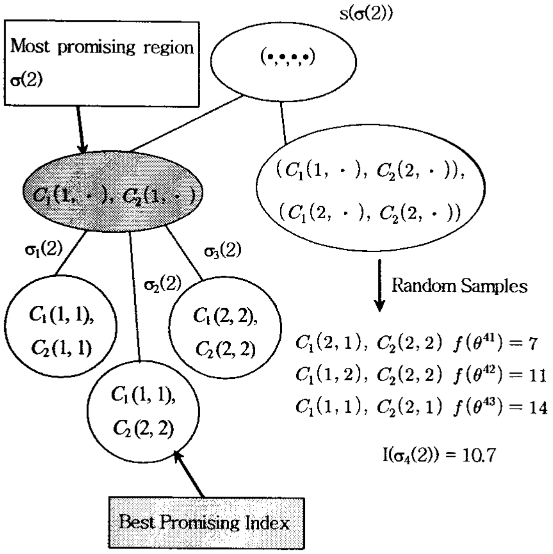
[Figure 1] Simple Example for Clustering using NP methodology

To simplify the problem, it is assumed each dimension has only two values. That is, $x_i = \{1, 2\}, i = 1, 2$. [Figure 2] and 3 demonstrates a partitioning tree where all features can take two different values and the problem is to find the optimal location of $m = 2$ clusters (identified as C_1

and C_2). This partitioning approach helps with the scalability of the method with respect to the feature dimension. It focuses on fixing one feature at a time and is in that sense monothetic, but not fully so as all features are randomly assigned values during the random sampling stage, and thus all features are used simultaneously to select subregions. This approach can thus be thought of as having elements of both monothetic and polythetic clustering.

It is also important to note that the partitioning tree imposes a structure on the space of all possible clusters, and thus determines the efficiency of the search through this space. Furthermore, investigating effective methods for ordering features is an important future research topic.

[Figure 2] show initial partitioning. In the first subset, 1st dimension of each cluster are set as (1, 1). In the second subset, 1st dimension of each cluster are set as (1, 2). In the third subset, 1st dimension of each cluster are set as (2, 2).

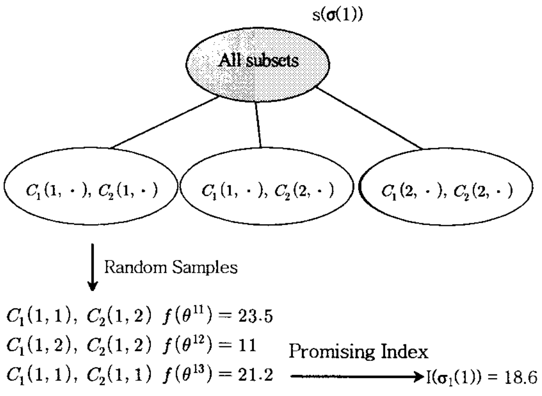

s(σ(2))
Most promising region σ(2)
(•,•,•,•)
$C_1(1, \cdot), C_2(1, \cdot)$
$(C_1(1, \cdot), C_2(2, \cdot))$, $(C_1(2, \cdot), C_2(2, \cdot))$
$\sigma_1(2)$
$\sigma_2(2)$
$\sigma_3(2)$
$C_1(1, 1)$, $C_2(1, 1)$
$C_1(2, 2)$, $C_2(2, 2)$
$C_1(1, 1)$, $C_2(2, 2)$
Random Samples
$C_1(2,1),\ C_2(2,2)\ f(\theta^{41}) = 7$
$C_1(1,2),\ C_2(2,2)\ f(\theta^{42}) = 11$
$C_1(1,1),\ C_2(2,1)\ f(\theta^{43}) = 14$
$I(\sigma_4(2)) = 10.7$
Best Promising Index

[Figure 2] 1st iteration of the example

After partitioning three subsets, random sampling is applied. For example, for the first subset, every sample point of this subset has a fixed first dimension; the first cluster and the second cluster are fixed as 1. For the remaining dimension, centers can be randomly assigned from the values {1, 2}. Using the sampling from each subset, a similarity value is calculated by formulation (2).

Based on these values, the promising index is calculated for each subset. After calculating promising index for all subsets, the most promising region is the 1st subset because it has the smallest promising index.

[Figure 3] shows that 1st iteration and the most promising index is 1st subset. The partitioning of the 2nd iteration starts from the 1st subset. Because the 2nd dimension can take 2 different values, three different subsets can be obtained like in the 1st iteration. Also, the 2nd iteration is the


s(σ(1))
All subsets
$C_1(1, \cdot), C_2(1, \cdot)$
$C_1(1, \cdot), C_2(2, \cdot)$
$C_1(2, \cdot), C_2(2, \cdot)$
Random Samples
$C_1(1,1),\ C_2(1,2)\ f(\theta^{11}) = 23.5$
$C_1(1,2),\ C_2(1,2)\ f(\theta^{12}) = 11$
$C_1(1,1),\ C_2(1,1)\ f(\theta^{13}) = 21.2$
Promising Index
$I(\sigma_1(1)) = 18.6$

[Figure 3] 2nd iteration of the example

maximum depth because there are two dimensions in this problem. From the 2nd iteration, there is one more subset which is called the surrounding region. The subset which contains center $(C_1(1,\bullet), C_2(2,\bullet))$ in [Figure 3] is surrounding region. After sampling from all subsets, the most promising index is found in the second region, having the 1st cluster's coordinate (1, 1) and the 2nd cluster's coordinate (1, 2). These coordinates are optimal that minimize the similarity of the problem.

3.4 NPCLUSTER Algorithm

NPCLUSTER Algorithm is modified algorithm for data clustering from NP method which guarantees to global convergence (Shi and Olafsson, 2000). So, NPCLUSTER also converges to the global optimum.

- **NPCLUSTER Algorithm**

Step 1 : *Initialization*

Set $k=0$ and $\sigma(k)=\Theta$.

Step 2 : *Partitioning*

If $d(\sigma(k)) \neq d^*$, that is $\sigma(k) \neq \Sigma_0$, partition the most promising region $\sigma(k)$, into $M_{\sigma(k)}$ sub-regions $\sigma_1(k), \cdots, \sigma_M(k)$, and aggregate the surrounding region $\Theta \setminus \sigma(k)$ into one region $\sigma_{M_{\sigma(k)}+1}(k)$.

if $d(\sigma(k)) \neq d^*$, then let $M_{\sigma(k)}=1$ and $\sigma_1(k)=\sigma(k)$.

Step 3 : *Random Sampling*

Step 3-1 : Let $i=1$.

Step 3-2 : K-means Algorithm

Assign instances using random sampling to the cluster center and calculate performance value which is mentioned in (2). Repeat this, until the centers no longer move for each of the regions,

$\sigma_j(k),\ j=1,2,\cdots,M_{\sigma(k)}+1$

Step 3-3 : If $i=n_0$, continue to Step 4. Otherwise, let $i=i+1$ and go back to Step 3~2. Through the K-means Algorithm, from each of the region, repeatedly obtain n_0 sample sets, that is θ^{j1}, $\theta^{j2}, \cdots, \theta^{jn_0}$, from each of the regions, $\sigma_j(k), j=1,2,\cdots,M_{\sigma(k)}+1$ according to the distribution, and calculate the corresponding performance values which is mentioned in (2)

$f(\theta^{j1}),\ f(\theta^{j2}),\cdots, f(\theta^{jn_0})$.

Step 4 : *Estimating the Promising Index*

Let the overall sample mean be the promising index for all subregions,

$j=1,2,\cdots,M_{\sigma(k)}+1$,

$\hat{I}(\sigma_j(k)) = \overline{f_j}(k) = \Sigma_{i=1}^{n_0} f(\theta^{ji}(k))/n_0$

Step 5 : *Backtracking*

Select the index of the region with the best promising index,

$\hat{j}_k \in \arg \min_{j\in\{1,\cdots,M_{\sigma(k)}+1\}} \hat{I}(\sigma_j)$

for all $j=1,2,\cdots,M_{\sigma(k)}+1$.

If more than one region is equally promising, the tie can be broken arbitrarily. If this index corresponds to a region that is a sub-region, $\sigma(k)$, then let this be the most promising region in the next iterations. Otherwise, if the index corresponds to the surrounding region, backtrack to a larger region containing the current most promising region. That is, let

$$\sigma(k+1)=\begin{cases}\sigma_{\hat{j}_k}(k), & \text{if } \hat{j}_k < M+1\\ s(\sigma(k)), & otherwise\end{cases}$$

Step 6 : *Checking Stopping Rule*

If $\sigma(k+1)\in\Sigma_0$, stop and let $\sigma_{opt}=\sigma(k+1)$ else $k=k+1$ and go back to Step 2.

3.5 Numerical Evaluation

To evaluate the effectiveness of the NPCLUSTER, we compare it with the PAM algorithm, which is a variant of the k-medoids approach (Kaufman and Rousseeuw, 1990), and CLARA, its more scalable variation.

The motivation for the selection of these algorithms for comparison is that like NPCLUSTER, these algorithms use a partitional approach to identify cluster centers and employ a random sampling strategy to improve scalability.

We use three realistic data sets from the UCI repository of machine learning databases (Blake and Merz, 1998). The characteristics of these sets are shown in <Table 1>, from which we note that the sizes ranges from 286 to 958 instances and from 9 to 10 features. We ran 60 replications for each experiment and report both similarity value with average and standard error and computation time.

<Table 1> Characteristics of the Tested Data Sets

Data Set	Instances	Features
Breast cancer	286	9
Wisconsin Breast Cancer	699	10
Tic-Tac-Toe	958	9

The quantity of the clusters. ie. computation time is used complexity, real calculation number. And the quality of the clusters obtained is measured by the compactness or similarity value of the clusters which is defined in (2). The smaller this value is, the better the clustering becomes. We note that although PAM has the best performance in terms of similarity values, it comes at a very high computation cost. By using sampling in case of NPCLUSTER and CLARA, the computation time can be reduced by two orders of magnitude. The CLARA algorithm, on the other hand, uses the least computation time for three data sets but the quality of the clusters is not satisfactory compared to the other methods.

The ability of the NPCLUSTER method to

<Table 2> Comparison of Algorithms for Similarity Value and Computation Time

Data Set	NPCLUSTER		PAM		CLARA	
	Similarity Value	Computation Time	Similarity Value	Computation Time	Similarity Value	Computation Time
Breast Cancer	1302.3 ± 13	$0.8 \cdot 10^5$	978 ± 2	$19.2 \cdot 10^5$	1971 ± 21	$0.8 \cdot 10^5$
Wisconsin Breast Cancer	4259.0 ± 46	$3.9 \cdot 10^5$	3166 ± 0.5	$102.6 \cdot 10^5$	5026 ± 54	$1.7 \cdot 10^5$
Tic-Tac-Toe	5130 ± 7.2	$7.4 \cdot 10^5$	4025 ± 5.7	$194.3 \cdot 10^5$	7106 ± 64	$1.8 \cdot 10^5$

use sampling and still obtain high quality solution stems from the fact that when an incorrect move is made in the partitioning tree, it can be corrected in the next (or a later) iteration, when a new sample of instances indicates that this was the wrong move. Thus, if there is a large amount of noise in the performance estimates (i.e., a small sample of instance is used), then the algorithm may backtrack frequently. Frequent backtracking implies more iterations, and thus increases computation time so there is a tradeoff between fast computation in each iteration and more iterations needed when reducing the instance sample size. However, we note that the NPCLUSTER algorithm achieves this balance in an automated manner.

<Table 3> Numerical Results for Different Percentage of instances Used

Data Set	Fraction	Similarity Value	Computation Time	Backtracking
		Avg. ± S.E.	Avg. ± S.E.	Avg. ± S.E.
Breast Cancer	100%	1302.3 ± 13	84115 ± 3166	0.72 ± 0.05
	50%	1276.6 ± 15	27295 ± 901	0.30 ± 0.07
	25%	1322.9 ± 12	25699 ± 689	0.56 ± 0.19
	5%	1363.1 ± 13	27698 ± 960	0.44 ± 0.12
	0.5%	1430.9 ± 12	33773 ± 1203	0.34 ± 0.08
Wisconsin Breast Cancer	100%	4259.0 ± 46	394777 ± 6668	0.14 ± 0.05
	50%	4207.7 ± 53	101666 ± 1352	0.08 ± 0.04
	25%	4264.7 ± 51	43794 ± 498	0.08 ± 0.04
	5%	4363.4 ± 41	38966 ± 590	0.10 ± 0.05
	0.5%	4401.1 ± 49	44065 ± 775	0.14 ± 0.06
Tic-Tac-Toe	100%	5130 ± 7.2	745131 ± 4935	0.00 ± 0.00
	50%	5278 ± 8.4	191796 ± 1718	0.00 ± 0.00
	25%	5267 ± 11.3	59501 ± 443	0.00 ± 0.00
	5%	5351 ± 16.6	14514 ± 21	0.00 ± 0.00
	0.5%	5812 ± 46.0	14975 ± 20	0.00 ± 0.00

4. Numerical Results of Instance subset

As has been noted above, repeated calculation of cluster performance according to (2) is time consuming and a more scalable approach is to use an estimate

$$\hat{f}(x^{(1)}, x^{(2)}, \cdots, x^{(k)}, I) \qquad (3)$$

that is calculated from a (small) subset I of the set of all instances. The key questions to be answered is how much savings in computation time can be achieved by using this estimate, what is the best sample size $|I|$, and how sensitive the clustering algorithm performance is to this sample size.

To obtain some tentative answers to these questions and to demonstrate feasibility for the scalability improvements that are possible by using sampling, we apply the NPCLUSTER method described above to the same three data sets as before. The numerical results are reported in <Table 3>, which shows the solution quality (similarity value), computation time, and average amount of backtracks for varying amounts of sampling. These results clearly indicate that random sampling can be effectively used to achieve substantial computational benefits without sacrificing solution quality. We recall that here solution quality is defined by equation (2) as being a measure of within cluster similarity that is the sum of the deviation of instances from the cluster center of the cluster to which they are assigned.

For example, using 25% of the Breast

Cancel data set reduces the computation time by 69% while the similarity only increases by 2% and is within two standard deviations of the value without sampling. Similarly for the Breast Cancel data set, by using 5% of the instances in each step, computational time can be reduced, by 90% while similarity value is only increased by 2 percent. Similar results can get from Tic-Tac-Toe data set. Computation time reduction is 98% while similarity value increase is 4%.

Furthermore, the performance is not very sensitive to exact selection of an amount of sampling. For example, for both problems (Breast Cancer and Wisconsin Breast Cancer) the performance when 5% of instances is used is very similar to the performance when 25% of instances is used. In case of Tic-Tac-Toe, more larger data set, same thing happens between 5% of instances and 0.5% of instances. Thus, the iterative nature and automatic backtracking feature of the NPCLUSTER algorithm allow us to achieve significant computational improvements without exact calibration of how many instances are needed. And this is more effective in case of large data set

We also note that the variability of the performance (as measured by the standard error reported in <Table 3>) is stable. This is somewhat surprising as one might expect that dealing with the noisier sets corresponding to a small sample of the original data might give rise to higher variability. The fact that such an increase is not observed is an indication that the NPCLUSTER is very effective in dealing with such uncertainty.

There is, on the other hand, a significant observed change in the variability of the computing time. For both test problems (Breast Cancer and Wisconsin Breast Cancer), the computation time is the least for when using 25% of the database. The reduction in variability is not, however, solely explained by the shorter computation time as the estimated coefficient of variation (standard error divided by the average) shows a similar pattern. For example, for the Wisconsin Breast Cancer data the estimated coefficient of variation is shown in <Table 4>. Thus, we conclude that sampling does not only reduce the computation time, but the computation time is more stable when the sample size is selected appropriately.

<Table 4> Estimated Coefficient of variation

Fraction	CV
100%	$1.69 \cdot 10^2$
50%	$1.33 \cdot 10^2$
25%	$1.14 \cdot 10^2$
5%	$1.51 \cdot 10^2$
0.5%	$1.76 \cdot 10^2$

5. Conclusions and Future Research

Clustering is one of the most important areas of knowledge discovery in databases, and the use of optimization techniques for clustering offers considerable promise. The scalability of such techniques is one of the key issues to be addressed as the field progresses. In particular, scalability with respect to increasing number of instances is critical as databases become ever larger. One way of dealing with this issue is to use a subset of all instances for the learning algorithm. The obvious tradeoff is between computational issues, where fewer instances imply faster learning, and solution quality, where using fewer instances may imply lower quality models.

We have designed an optimization bases approach to the partitional clustering problem where the algorithm is specifically designed to deal with noisy performance estimates, such as those that arise when only part of the data is used to create clusters. Numerical results show that considerable speedup can be achieved (up to 90% for the numerical examples) with no or minimal reduction in solution quality. Also, the algorithm is robust with respect to the amount of instances used so there is no need to carefully determine the fraction of the database that needs to be used.

There are numerous issues that should be addressed for further development of this methodology. For example, an extensive numerical evaluation on a variety of realistic and synthetic problems should be performed, relating the computational speedup to characteristics of the data, and developing heuristic for specifying amount of instances to be used and evaluating their robustness. Especially, large data should be evaluated to robust scalability.

As noted in Section 3.3 above, determining intelligent ways of ordering the features is also a critical issue. The partitioning tree imposes a structure on the search space of all possible clusters and the order in which features are considered determines this structure. Thus, an important future research topic is to investigate how the algorithm can be improved by determining a generic way of creating high quality partitioning for arbitrary clustering problems.

Finally, the ability of this approach to handle arbitrary performance functions opens up some interesting possibilities. In this paper we only considered a measure of cluster compactness, but as noted in Section 3.2, any measure can be used. Thus, it is of interest to apply the new algorithm with various measures of cluster performance and compare qualities of the resulting cluster. In other words, by using the same optimization methodology, but different measures of what makes a good cluster, and analyzing the resulting clusters, we believe insights into data clustering in general could be obtained.

References

[1] Basu, A., "Perspectives on operations research in data and knowledge management", *European Journal of Operational Research*, Vol.111(1998), 1~14.

[2] Blake, C. L. and C. J. Merz, UCI Repository of Machine Learning Databases [http://www.ici.uci.edu/mlearn/MLRepository.html]. Department of Information and Computer Science, University of California, Irvine, CA (1998).

[3] Bradley, P., U. Fayyad, and C. Reina, "Scaling clustering algorithms to large databases", *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining,* (1998), 9~15.

[4] Bradley, P., J. Gehrke, R. Ramakrishnan, and R. Srikant, "Scaling mining algorithms to

large databases", *Communications of the ACM*, Vol.45, No.8(2002), 38~43.

[5] Bradley, P. S., Mangasarian, O. L., and Street, W. N., "Feature selection via mathematical programming", *INFORMS Journal on Computing*, Vol.10(1998), 209~217.

[6] Chauchat, J-H. and R. Rakotomalala, "Sampling strategies for building decision trees from very large databases comprising many continuous attributes", In Liu and Motada (eds.) Instance Selection and Construction for Data Mining, Kluwer, (2001).

[7] Domingo, C. Gavalda R., and Watanabe, R., "Adaptive Sampling Methods for Scaling Up Knowledge discovery", *Data Mining and Knowledge Discovery*, Vol.6, No.2(2002), 131~152.

[8] Estevill-Castro, V., "Why so many clustering algorithms", *SIGKDD Explorations*, Vol.4, No.1(2002), 65~75.

[9] Ester, M., H.-P. Kriegel, and X. Xu, "Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification", *Proceedings of the 4th International Symposium of Large Spatial Databases* (1995), 67~82.

[10] Farnstrom, F., J. Lewis, and C. Elkan, "Scalability for clustering algorithms revisited", *SIGKDD Explorations*, Vol.2, No.1 (2000), 51~57.

[11] Forman, G. and B. Zhang, "Distributed data clustering can be efficient and exact", *SIGKDD Explorations*, Vol.2, No.2(2000), 34~38.

[12] Guha, S. R. Rastogi, and K Shim, "CURE: An efficient clustering algorithm for large databases", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, (1998), 73~84.

[13] Grabmeier, J. and A. Rudolph, "Techniques of cluster algorithms in data mining", *Data Mining and Knowledge Discovery*, Vol.6 (2002), 303~360.

[14] Jain, A. K., Murty, M. N. and Flynn, P. J., "Data clustering : a review", *ACM Computing Surveys*, Vol.31(1999), 264~323.

[15] John, G. and P. Langley, "Static versus dynamic sampling for data mining", *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1996), 367~370.

[16] Kaufman, L. and P. J. Rousseeuw, *Finding Groups in Data : An Introduction to Cluster Analysis*. John Wiley & Sons, New York (1990).

[17] Kim, J. M., "Optimization under uncertainty with application to data clustering", Ph.D. Diss., Dept. of IMSE, Iowa State University (2002).

[18] Kiven, J. and H. Mannila, "The power of sampling in knowledge discovery", *ACM Symposium on Principles of Database Theory* (1994), 77~85.

[19] Liu, H. and H. Motoda, *Instance Selection and Construction for Data Mining*, Kluwer (2001).

[20] Ng, R. T. and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining", *Proceedings of 20th International Conference on Very Large Data Bases* (1994).

[21] Ólafsson, S., "Iterative ranking-and-selection for large-scale optimization", *Proceedings of the Winter Simulation Conference* (1999), 479~485.

[22] Ólafsson, S., "*Improving scalability of e-commerce systems with knowledge discovery*", In Prabu, Kumara and Kamath (eds.) Scalable Enterprise System-An Introduction to Recent Advances, Kluwer, (2003).

[23] Provost, F., D. Jenson, and T. Oates, "Efficient progressive sampling", *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining* (1999), 23~32.

[24] Provost, F. and V. Kolluri, "A survey of methods for scaling up inductive algorithms", *Data Mining and Knowledge Discovery*, Vol.3(1999), 131～169.

[25] Shi, L. and S. Ólafsson, "Nested partitions method for global optimization", *Operations Research*, Vol.48(2000), 390～407.

[26] Toivonen, H., "Sampling large databases for association rules", *Proceedings of the 22nd International Conference on Very Large Databases* (1996), 134～145.

[27] Yim D. S. and H. S. Oh, "Application of Genetic and Local Optimization Algorithms for Object Clustering Problem with Similarity Coefficients", *Journal of the Korean Institute of Industrial Engineers*, Vol.29, No.1(2003), 90～99.

[28] Zhang, T., R. Ramakrishnan, and M. Livny, "BIRCH : An efficient data clustering method for very large databases", *Proceedings of the ACM SIGMOD International Conference on Management of Data* (1996), 103～114.

요약

최적화에 기반 한 데이터 클러스터링 알고리즘

김주미*

대용량의 데이터 처리에 관한 문제는 데이터 마이닝 내 중요한 이슈 중의 하나이다. 특히 데이터 클러스터링과 같이 컴퓨터 시뮬레이션으로 인한 부하가 큰 경우 더더욱 그러하다. 그러나 대개 이러한 문제는 Random sampling 으로 어느 정도 해결이 가능하다. 문제는 이런 샘플링을 통해서 발생하는 noise의 해결이다. 본 논문에서는 그러한 noise 문제를 극복할 수 있도록 설계된 새로운 데이터클러스터링 알고리즘을 소개한다. 기존의 데이터 클러스팅 알고리즘과의 컴퓨터 비교 실험을 통해 본 알고리즘의 우수성을 밝혔으며 아울러 더 나아가 데이터 set의 일부만을 사용한 시뮬레이션 결과를 통해, 해의 정확도와 상관없이 실험 시간 또한 단축되었음을 보여주고 있다.

* 중소기업연구원, 연구위원