# Mutagenicity Risk Analysis by Using Class Association Rules

Takashi Washio, Koutarou Nakanishi, Hiroshi Motoda[1], and Takashi Okada[2]

[1] I.S.I.R., Osaka University
washio@sanken.osaka-u.ac.jp
[2] Kwansei Gakuin University
okada-office@ksc.kwansei.ac.jp

**Abstract.** Mutagenicity analysis of chemical compounds is crucial for the cause investigation of our modern diseases including cancers. For the analysis, accurate and comprehensive classification of the mutagenicity is strongly needed. Especially, use of appropriate features of the chemical compounds plays a key role for the interpretability of the classification results. In this paper, a classification approach named "*Levelwise Subspace Clustering based Classification by Aggregating Emerging Patterns* (LSC-CAEP)" which is known to be accurate and provides interpretable rules is applied to a mutagenicity data set. Promising results of the analysis are shown through a demonstration.

## 1 Introduction

Mutagenicity is one of the important biological activities of chemical compounds for our health [1]. Mutation is a structural alteration in DNA. In most cases, such mutations harm human health. A high correlation is also observed between mutagenicity and carcinogenicity. However, the experimental identification of the mutagenicity for all compounds is very hard due to the required cost and time of the experiments. Accordingly, an analytical screening of the mutagenic chemical compounds have been attempted in the field of Structure Activity Relationship (SAR) analysis [2]. The study of SAR between chemical structures and biological activity is a well-established domain in medicinal science. However, the automated classification of the mutagenicity based on the chemical structures is still difficult due to the variety of the structure activity relationships. In recent innovation of chemical analysis technologies, the high throughput screening on the biological activity of chemical compounds became to provide the vast amount of SAR data. The introduction of data mining techniques to SAR analysis will be an efficient remedy to the difficulty of the mutagenicity classification. The extensive and detailed analysis of SAR data by some new classification technique is expected to figure out chemical substructures causing the mutagenicity and to facilitate the drug development process.

In the field of data mining, a new rule-based classification framework has been proposed recently where each classification rule relates a class of instances with

T. Washio et al. (Eds.): JSAI 2005 Workshops, LNAI 4012, pp. 436–445, 2006.

a quantitative frequent itemset appearing in the class instances. This framework called "*Levelwise Subspace Clustering based Classification by Aggregating Emerging Patterns* (LSC-CAEP)" [3] is based on both of "*Subspace Clustering* (SC)" and "*Class Association Rule* (CAR)" techniques. It is known to have significant advantages on both accuracy and interpretability of the classification results. These advantages are highly desired for the SAR analysis on the mutagenicity where the feasible relationships between the classification results and the chemical substructures must be understood by chemists.

In this paper, some related work to LSC-CAEP is described at first to clarify its features. Second, the principle of CAR mining used in LSC-CAEP is explained. Third, the principle and the algorithm to mine QFIs which are the conditional parts of CARs of LSC-CAEP are outlined. Subsequently, the performance of LSC-CAEP is demonstrated for UCI benchmark data in comparison with some major classification approaches. Finally, the mutagenicity analysis of chemical compounds by using LSC-CAEP is shown together with the discussion on the analysis result by an expert chemist.

## 2 Related Work

Subspace Clustering (SC) is to search numeric attribute subspaces to obtain better clusters than these in the original attribute space. An initiative study CLIQUE performs an axis-parallel grid based clustering where maximal sets of connected dense blocks are searched by greedy merging the blocks in every subspace [4]. DOC seeks dense clusters in every subspace by counting the instances in axis-parallel windows [5]. The computational complexity of the representative grid and window based SC approaches is between $O(N)$ and $O(N \log N)$ which is tractable, where $N$ is the number of instances in data. However, they miss some clusters due to inadequate orientation, shape and size of the axis-parallel grids and windows. The recently developed SUBCLU searches density-based subspace clusters under a rigid density measure proposed by DBSCAN [6, 7]. It uses (anti-) monotonicity property for the dense clusters that the instances in a cluster in an attribute space are always included in some clusters in its subspaces. By combining this property with the Apriori algorithm, SUBCLU exhaustively derives all dense clusters in every attribute subspace. However, because it basically needs the pairwise distances among instances, its computational complexity under a well designed algorithm lies between $O(N \log N)$ and $O(N^2)$ which is often unacceptable for large data sets [8]. Another drawback of these approaches is the less interpretability, because the clusters having high dimensional and complex shapes are hardly understood.

To overcome this interpretability issue, the approaches to mine "*Quantitative Frequent Itemsets* (QFIs)" and "*Quantitative Association Rules* (QARs)" have been studied. The items including numeric interval attributes such as "$< Age : [32, 35] >$" are called "*numeric items.*" A numeric item having a unique numeric value is represented by using a point interval such as "$< NumCars : [2, 2] >$." On the other hand, the items including categorical attributes such as

"$< Married : Yes >$" are called "*categorical items*." An itemset consists of numeric and categorical items where each attribute does not appear in more than one item in the itemset, *i.e.*, any item does not share its attribute with the other items. An itemset is a QFI if it is supported by given instances (transactions) frequently more than a threshold called minimum support *minsup*. An example QFI is "$\{< Age : [30, 39] >, < Married : Yes >, < NumCars : [2, 2] >\}$" which states "There are many persons who are in their thirties, married and has two cars." Most of studies in this field take preprocessing approaches to partition the value range of each numeric attribute into some intervals and to greedily merge the adjacent intervals [9, 10]. The conventional Basket Analysis is subsequently applied to find QFIs. Their complexity is $O(N \log N)$. QFIs and QARs are comprehensive to analyze the clusters and their inter-subspace relations. However, the discretization of the entire numeric attribute space is not optimal for local instance distribution in each subspace, and the greedy merging may miss the optimal discretization.

The recent study is extending to classification based on "*Class Association Rules* (CARs)" where the body is a QFI and the head a class value. Given a training data set $D$ which is an attribute-value and class table or a set of class labeled transactions, let $D_{cl}$ be a set of all instances having a class $cl$ in $D$. The body of a CAR is a QFI which is supported by $D_{cl}$ more frequently than a *minsup* threshold. The classification of an instance is made by using the CARs which bodies are included in the instance. CBA is an initiative work on this topic [11]. It seeks QFIs similarly to the above QFI mining, and subsequently CARs are searched. CMAR and CAEP are the successors to improve the performance by using multiple CARs for a classification [12, 13]. Especially the CAEP shows better performance in comparison with the conventional classifiers such as C4.5. However, these approaches have a problem on the optimality of the discretization of the entire numeric attribute space similarly to the above QAR mining approaches.

Aforementioned LSC-CAEP is an efficient remedy to overcome these difficulties. It is based on a novel Subspace Clustering (SC) technique similar to SUBCLU which is based on a rigid density measure proposed by DBSCAN while extending the algorithm to process both numeric and categorical items. This technique enables a complete mining of QFIs while reducing the computational complexity to $O(NlogN)$ by the application of the density measure to each numeric attribute axis. LSC-CAEP derives accurate and comprehensive Class Association Rules (CARs) together with QFIs consisting of categorical items and an axis-parallel and hyper-rectangular and optimum cluster in a numeric attribute subspace.

## 3 Principle of CAR Mining

CARs in LSC-CAEP are derived by following the principle of CAEP [13]. The training of CAEP consists of two processes. The first process is to derive all rule bodies of CARs. Let the support of an itemset $a$ by $D_{cl}$ be $support_{D_{cl}}(a) = |\{t \in$

$D_{cl}|a \in t\}|/|D_{cl}|$. For every $cl$, a set of QFIs, $LQFI(cl)$, in which every itemset $a$ satisfies $support_{D_{cl}}(a) \geq minsup$, is derived from $D_{cl}$. In the original CAEP, this is done through a procedure identical to the standard Apriori algorithm whereas a SC is applied in this paper as described later. Subsequently, for every $a \in LQFI(cl)$, the following "*growth rate*" of $a$ for a class $cl$ is calculated. Let $\bar{D}_{cl} = D - D_{cl}$ be the opponent instances of $cl$.

**Growth rate**
If $support_{\bar{D}_{cl}}(a) \neq 0$, $growth\_rate_{\bar{D}_{cl} \rightarrow D_{cl}}(a) = \frac{support_{D_{cl}}(a)}{support_{\bar{D}_{cl}}(a)}$,
If $support_{\bar{D}_{cl}}(a) = 0$ and $support_{D_{cl}}(a) \neq 0$, $growth\_rate_{\bar{D}_{cl} \rightarrow D_{cl}}(a) = \infty$,
Otherwise $growth\_rate_{\bar{D}_{cl} \rightarrow D_{cl}}(a) = 0$.

When the growth rate $a$ is more than a "*growth rate threshold*" $\rho(> 1)$, *i.e.*, $growth\_rate_{\bar{D}_{cl} \rightarrow D_{cl}}(a) \geq \rho$, $a$ is selected as a rule body where its head has the class $cl$, *i.e.*, $a \Rightarrow cl$. The underlying principle here is to select the rule bodies having the strength to differentiate the class $cl$ from the others. This is more advantageous than the confidence based rule selection of CBA and CMAR. Even if the rule confidence is high in $D_{cl}$, the rule can match many instances in $\bar{D}_{cl}$. Such rules are weak for classification.

The second process is to derive a "*base score*" of each $cl$ which is a weighting factor on the votes for class prediction. First, a strength of a rule body $a$ is introduced as $support_{D_{cl}}(a)/(support_{D_{cl}}(a) + support_{\bar{D}_{cl}}(a)) = growth\_rate_{\bar{D}_{cl} \rightarrow D_{cl}}(a)/(growth\_rate_{\bar{D}_{cl} \rightarrow D_{cl}}(a) + 1)$. This is because the rule strength is mainly defined by the relative difference between $support_{D_{cl}}(a)$ and $support_{\bar{D}_{cl}}(a)$. Let $LRB(cl)$ be the set of all rule bodies selected from $LQFI(cl)$ in the aforementioned process. The following "*aggregate score*" of an instance $t$ for a class $cl$ represents the possibility of $t$ to be classified into $cl$ by the rule bodies in $LRB(cl)$.

**Aggregate score**

$$score(t, cl) = \sum_{a \subseteq t, a \in LRB(cl)} \frac{growth\_rate(a)}{growth\_rate(a) + 1} * support_{D_{cl}}(a). \qquad (1)$$

Because the number of rule bodies in $LRB(cl)$ may not be balanced among classes, instances usually may get higher scores for some specific classes. To eliminate this bias, the base score is introduced to weight each class $cl$.

**Base score:**
$base\_score(cl)$ is the aggregate score where the number of instances having their aggregate scores less than this score is $Tail\%$ of all instances in $D_{cl}$.

The classification of CAEP is performed based on the CARs obtained in the training phase. It uses the results of $base\_score(cl)$, $growth\_rate(a)$ and $support_{D_{cl}}(a)$ for all classes $cl$ and all $a \in LRB(cl)$ obtained in the training phase. Given an instance $t$ to be classified, its aggregate score for $cl$, $score(t, cl)$, is computed from these results and Eq.(1). Then, it is normalized by $base\_score(cl)$ to eliminate the aforementioned bias as follows.

**Normalized score**
$norm\_score(t, cl) = \frac{score(t,cl)}{base\_score(cl)}$.

$cl$ having the maximum normalized score is assigned to the class of $t$. Except the derivation of $LQFI(cl)$ for all $cl$, the computational complexity of the training and the classification is $O(N)$ where $N = |D|$, since it scans the training data only twice.

## 4 Mining Rule Bodies of CARs

LSC-CAEP searches QFIs of rule bodies from a data set $D_{cl}$ where each transaction consists of numeric and categorical items. LSC-CAEP assumes that dense clusters of the transactions exist with scattered background noise in the subspace. The upper part of Fig. 1 depicts this example where every numeric item takes a point interval (unique) value in each transaction, and two dense clusters exist in a two dimensional attribute subspace $S = \{p_1, p_2\}$.

LSC-CAEP uses a definition of density similar to DBSCAN. This approach significantly reduces the possibility to miss clusters under an appropriate density threshold. LSC-CAEP uses a levelwise algorithm where it starts from the clusters in one dimensional subspaces, and joins $(k-1)$ dimensional clusters into a candidate cluster $\hat{C}^S$ in $k$ dimensional subspace $S$. While this is similar to SUBCLU, LSC-CAEP can derive clusters on both numeric and categorical items by embedding the levelwise subspace clustering into the standard Apriori algorithm. At each level, first, it derives frequent itemsets consisting of categorical items and numeric item's attributes, then second, dense clusters in $S$ formed by the numeric attributes in the frequent itemsets are searched. The clusters supported more than a minimum support ($minsup$) in numeric and categorical attribute subspaces are exhaustively mined.

To avoid $O(N^2)$ computational complexity, LSC-CAEP does not compute the pairwise distances among transactions. Instead, it projects transactions in a candidate dense cluster $\hat{C}^S$ onto each attribute axis of the subspace $S$. The upper part of Fig. 1 shows a case that $\hat{C}^S$ is a $[0, 100] \times [0, 100]$ region. All maximal density-connected sets are searched in the transactions projected onto every axis $p$, where a density-connected set on $p$ is such that for each transaction in the set $\pm\Delta_p$ neighborhood on $p$ has to contain at least a minimum number of $MinPts$ transactions, and a maximal density-connected set is a density-connected set which is not contained in any other density-connected set. An intersection of the maximal density-connected sets on all axes in the subspace becomes a new $\hat{C}^S$ due to the (anti-)monotonicity of the density. In Fig. 1, the four intersections are new $\hat{C}^S$. These projection and searching maximal density-connected sets are iterated until each $\hat{C}^S$ converges to a dense cluster $C^S$. The two intersections containing the dense clusters in Fig. 1 are retained under this iteration and the rest pruned. In the lower part of Fig. 1, dense region of each retained intersection is further narrowed down to ensure the density within the region projected to every axis. Because the density on every axis is evaluated in a scan of sorted transactions, the complexity of this algorithm is expected to be $O(N \log N)$.

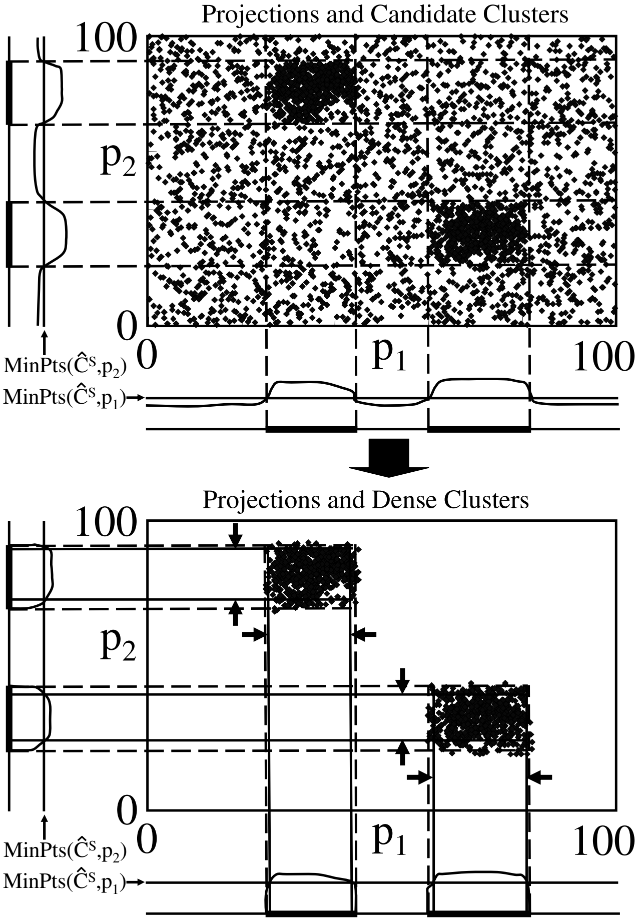

**Fig. 1.** Derivation of dense clusters

In the search of maximal density-connected sets on an axis, if $MinPts$ is lower than the background noise level, the projection of dense clusters may be buried in the background. If it is too high, the projection of dense clusters may be missed. Accordingly, $MinPts$ is adapted to $MinPts(\hat{C}^S, p)$ which is the expected number of transactions projected to the $\pm\Delta_p$ neighborhood on an axis $p$ from each $\hat{C}^S$ assuming that $\hat{C}^S$ has the average density of the subspace $S$. $MinPts(\hat{C}^S, p)$ is always between the densities of the dense cluster and the background. In Fig. 1, $MinPts(\hat{C}^S, p)$ efficiently extracts the maximal density-connected sets reflecting the dense clusters. This adaptive density threshold further accelerates LSC-CAEP, because $MinPts(\hat{C}^S, p)$ is higher for a lower subspace dimension, and prunes more maximal density-connected sets below the noise level. In summary, LSC-CAEP takes the input parameters $\Delta_p$ (usually given by a unique relative width $\alpha_\Delta$ over the total range of data on every axis) and $minsup$.

## 5 Evaluation of Classification Performance

Table 1 shows the comparison of accuracy among C4.5, CBA and LSC-CAEP in the experiments by using 23 data sets in UCI repository. The accuracies of C4.5 and CBA were evaluated through 10 fold cross validations. J48 implemented in a data mining tool, Weka [15], was used for C4.5. CBA was obtained from its authors. The default parameters are applied to C4.5 and CBA.

The optimal parameters of LSC-CAEP was determined by a grid search of the parameter combinations to minimize the average classification error of 10 fold cross validations for each data set. The parameters of LSC-CAEP are *minsup*, $\alpha_\Delta$, $\rho$ and *Tail*. Mining QFIs for the rule bodies which computational complexity is $O(NlogN)$ needs the parameters of *minsup* and $\alpha_\Delta$. The other part to mine the relations between the rule bodies and the rule heads by following the principle of CAEP is only $O(N)$, and takes $\rho$ and *Tail*. Accordingly, the grid search of $\rho$ and *Tail* was made over their wide ranges, whereas the search on *minsup* and $\alpha_\Delta$ was limited to their feasible ranges based on our experience.

**Table 1.** Comparison of accuracies

| dataset | num. of records | num. of attributes(numeric) | num. of classes | C4.5 | CBA | LSC-CAEP | SD of LSC-CAEP |
|---|---|---|---|---|---|---|---|
| Australian | 690 | 14(6) | 2 | .8608 | .8538 | **.8666** | .0347 |
| Cars | 392 | 7(6) | 3 | .9617 | .9744 | **1.0000** | 0 |
| Cleve | 303 | 13(5) | 2 | .7656 | .8283 | **.8383** | .0422 |
| Crx | 690 | 15(6) | 2 | .8608 | .8538 | **.8715** | .0442 |
| Diabetes | 768 | 8(8) | 2 | .7226 | **.7445** | .7229 | .0681 |
| Ecoli | 336 | 8(7) | 8 | **.8422** | .7018 | .7794 | .0992 |
| German | 1000 | 20(7) | 2 | .7070 | **.7350** | .7173 | .0517 |
| Heart | 270 | 13(6) | 2 | .7666 | 8187 | **.8222** | .0694 |
| Hepatitis | 155 | 19(6) | 2 | **.8387** | .8182 | .8236 | .1062 |
| Horse | 368 | 22(8) | 2 | .6933 | .8236 | **.8394** | .0488 |
| Hypo | 3163 | 25(7) | 2 | **.9889** | .9826 | .9793 | .0071 |
| Iris | 150 | 4(4) | 3 | .9600 | .9467 | **.9733** | .0466 |
| Labor | 57 | 16(8) | 2 | .7368 | .8633 | **.9500** | .1124 |
| Led7 | 3200 | 7(0) | 10 | .7337 | .7206 | **.7400** | .0117 |
| Lymph | 148 | 18(2) | 4 | .7635 |  | **.8157** | .1189 |
| Nursery | 12960 | 8(0) | 5 | **.9705** | .8289 | .9408 | .0048 |
| Pima | 768 | 8(8) | 2 | **.7382** | .7290 | .7141 | .0338 |
| Sonar | 208 | 60(60) | 2 | **.7884** | .7746 | .6681 | .1288 |
| Tae | 151 | 5(1) | 3 | **.5099** | .4717 | .5067 | .1470 |
| Tic-Toc-Toe | 958 | 9(0) | 2 | .8507 | .9959 | **1.0000** | 0 |
| waveform | 5000 | 21(21) | 3 | .7664 | **.7968** | .7886 | .0153 |
| Wine | 178 | 13(13) | 3 | .9382 | .9496 | **.9833** | .0374 |
| Zoo | 101 | 16(0) | 7 | .9207 | **.9709** | .9309 | .0477 |
| Average |  |  |  | .8123 | .8264 | **.8379** | .0555 |

The final classification accuracies of LSC-CEAP were evaluated through 10 fold cross validations over the randomly shuffled original data sets.

Table 1 indicates the top accuracies for each data set by a bold face. The bottom row shows the average accuracy of each classifier over the 23 data sets. The right most column shows the standard deviations of the accuracies of LSC-CAEP over the 10 fold cross validations. The difference between the best accuracy and the second best is smaller than the standard deviation for each data except car, nursery, tic-toc-toe. Accordingly, LSC-CAEP is not very significant in terms of the absolute difference of the accuracy from the other methods. However, the average accuracy of LSC-CAEP is higher than the other methods. Under a scoring to assign 2 points to the best method and 1 point to the second for each data, the total scores of C4.5, CBA and LSC-CAEP are 19, 18 and 32 points respectively. Moreover, LSC-CAEP took the first place for 12 data sets among 23. Under the assumption of equal accuracy of three methods, the probability of this fact which follows a binominal distribution $B(23, 1/3)$ is ${}_{23}C_{12}(1/3)^{12}(2/3)^{11} = 2.9\%$. Based on these observations, LSC-CAEP is concluded to outperform C4.5 and CBA.

## 6 Mutagenicity Risk Analysis

Data sets on the mutagenicity of 230 chemical compounds are released for a benchmark of KDD Challenge 2000 in PKDD2000 conference [16]. We applied the LSC-CAEP to a dataset called MOE.CSV which includes 2D descriptors generated using the MOE QuaSAR-Descriptors. This data contains 102 attributes which include weight, density, hydrophobicity, geometric and physical descriptors of each molecule such as diameters, surface areas, shape parameters, bond connectivity, numbers of atoms and bonds of each type, electric charge distribution parameters and van der Waals force parameters. The quantitative mutagenicity activity of each instance is discretized by an expert chemist into four class levels of Inactive, Low, Medium and High. The parameters of LSC-CAEP are set as $\alpha_{\Delta} = 0.1Cminsup = 0.01$, $\rho = 1.3$ and $Tail = 50\%$ according to parameter survey.

Figure 2 represents QFIs of each class on an der Waals force area and volume plain. They should have a positive correlation in physics, and this fact is clearly reflected. In addition, this result indicates that higher values of vdw.area and vdw.vol lead high mutagenicity. LSC-CAEP can easily discover this type of quantitative correlation and its association with class values among massive attributes while this has been difficult within the conventional statistics and data mining. The followings are a set of rules on the inactivity having significant aggregate scores.

$$\{< logP(o/w) : [1.69, -2.63] >\} \Rightarrow Inactive$$
$$\{< PEOE_PC+ : [0.659, 1.11] >, < PEOE_PC- : [-1.11, -0.659] >\} \Rightarrow Inactive$$
$$\{< radius : [3.0, -3.0] >, < vdw.area : [127.2, 162.4] >, < vdw.vol : [152.7, -198.1] >\} \Rightarrow Inactive$$

The expert chemist suggested based on these rules that LogP (hydrophobicity), vdw.area, vdw.vol, radius and PC (number of positive valence electrons) of each

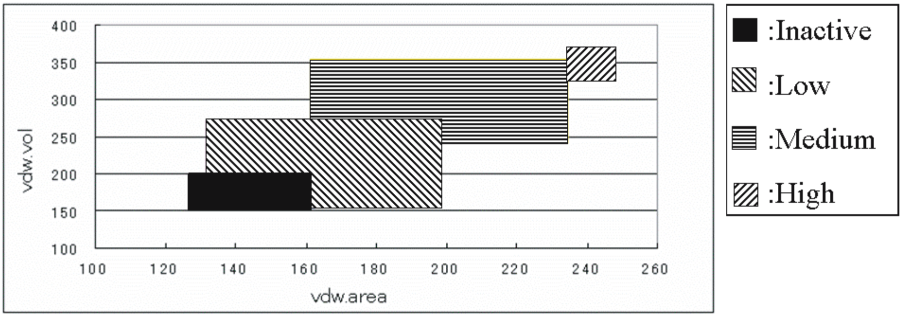

**Fig. 2.** QFIs of each class on vdw.area-vdw.vol plain

molecule are mutually correlated among inactive molecules, and could build a feasible assumption that hydrophobic and small molecules having less electric charge skewness have a tendency to be inactive. Based on the high rule interpretability, the risk of the mutagenicity of every chemical compound supported by chemical expertise can be predicted.

## 7 Conclusion

The generic high accuracy and interpretability of CARs derived by LSC-CAEP have been demonstrated through its application to the benchmark datasets of UCI and KDD2000 Challenge. Especially, its high practicality for chemical risk analysis has been demonstrated. Further study on the wide applicability of LSC-CAEP is currently underway.

## Acknowledgements

This study has been supported by Grant-in-Aid for Scientific Research No.16300045 and No.17650042 of the Japan Society for the Promotion of Science (JSPS).

## References

1. Debnath, A.K. et al.: Structure-Activity Relationship of Mutagenic Aromatic and Heteroaromatic Nitro compounds. J. Med. Chem. **34** (1991) 786–797
2. Klopman G.: Artificial Intelligence Approach to Structure-Activity Studies. J. Amer. Chem. Soc. **106** (1984) 7315–7321
3. Washio, T., Nakanishi, K., Motoda, H.: Deriving Class Association Rules Based on Levelwise Subspace Clustering. Proc. of PKDD2005: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases. LNAI **3721** (2005) 692–700

4. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications. Proc. of the 1998 ACM SIGMOD international conference on Management of data (1998) 94–105
5. Procopiuc, C.M., Jones, M., Agarwal, P.K., Murali, T.M.: A Monte Carlo algorithm for fast projective clustering. Proc. of the 2002 ACM SIGMOD international conference on Management of data. (2002) 418–427
6. Kailing, K., Kriegel, H.P., Kroger, P.: Density-Connected Subspace Clustering for High-Dimensional Data. Proc. Fourth SIAM International Conference on Data Mining (SDM'04). (2004) 246–257
7. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. (1996) 226–231
8. Brecheisen, S., Kriegel, H.P., Pfeifle, M.: Efficient density-based clustering of complex objects. Proc. of Fourth IEEE International Conference on Data Mining (2004) 43–50
9. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. Proc. of 1996 ACM SIGMOD Int. Conf. on Management of Data. (1996) 1–12
10. Wang, K., Hock, S., Tay, W., Liu, B.: Interestingness-based interval merger for numeric association rules. Proc. of 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD) (1998) 121–128
11. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. Proc. of Fourth International Conference on Knowledge Discovery and Data Mining (1998)
12. Li, W., Han, J., Pei, J.: Cmar: Accurate and efficient classification based on multiple class-association rules. Proc. of First IEEE International Conference on Data Mining (2001) 369–376
13. Dong, G., Zhang, X., Wong, L., Li, J.: Caep: Classification by aggregating emerging patterns. Proc. of Second International Conference on Discovery Science, LNCS **1721** (1999) 30–42
14. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning **29** (1997) 131–163
15. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques (2nd ed.). Morgan Kaufmann (2005).
16. Okada, T.: Guide to the Mutagenicity Data Set. KDD Challenge 2000 in PKDD2000: The Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases. (2000) http://www.clab.kwansei.ac.jp/mining/datasets/PAKDD2000/okd.htm