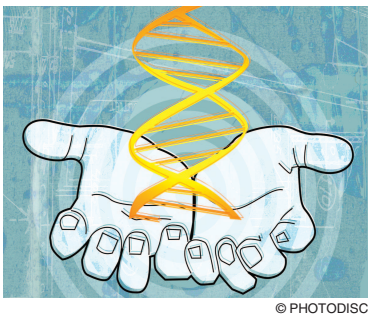
Chuan-Kang Ting, Wei-Ming Zeng, and Tzu-Chieh Lin, *National Chung Cheng University, TAIWAN*

Linkage Discovery Through Data Mining

1. Introduction

Genetic algorithms (GAs) are extensively adopted in various aspects of data mining, e.g., association rules, clustering, and classification [1, 2, 3]. Instead of applying GAs for data mining, this study addresses linkage discovery, an essential topic in GAs, by using data mining methods. Inspired by natural evolution, GAs utilize selection, crossover, and mutation operations to evolve candidate solutions into global optima [4]. This evolutionary scheme can effectively resolve many search and optimization problems. As the most salient feature of GAs, crossover enables the recombination of good parts of two selected chromosomes, yet, in doing so, may disrupt the collected promising segments.

Linkage discovery (or linkage identification) attempts to identify the tightly linked genes and bind them together to form building blocks, which can be used to avoid the disruptiveness induced by crossover operation [4, 5]. Identifying linkage is based on methods classified as estimation of distribution algorithms (EDAs), perturbation-based methods (PMs), and mixed methods. Rather than manipulating linked genes deterministically, EDAs attempt to derive probability models from selected chromosomes. Related methods include Bayesian optimization algorithm (BOA) [6], factorized distribution algorithm (FDA) [7], and extended compact genetic algorithm (ECGA) [8]. While considering linkage in a statistical manner, EDAs generally tend to neglect building blocks with a relatively low fitness contribution. Alternatively, PMs perturb certain genes of each chromosome in the population and, then, identify the linkage relationship by using the influence of perturbation. Related methods include linkage identification by nonlinearity check (LINC) [9, 10], linkage identification by non-monotonicity detection (LIMD) [11], linkage identification based on epistasis measures (LIEM) [12], and their variants. Although capable of detecting the linkage with low-contribution building blocks, PMs require additional fitness evaluations in perturbation. As another alternative, a representative of mixed methods is the dependency detection for distribution derived from fitness differences (D^5) [13]. This method extracts information from subsequently perturbing genes to estimate the linkage relationship. Tsuji and Munetomo [14] indicated that D^5 possesses the advantages of EDAs in terms of efficiency and of PMs in terms of detecting low-contribution building blocks. A more detailed survey of linkage discovery methods can be found in [14, 15].



© PHOTODISC

Rather than deriving the linkage relationship, this study proposes mining the interactions between genes for linkage. Figure 1 illustrates the metaphor that connects linkage discovery to data mining. Learning association rules involves finding the implicitly associated items from transactions, which is analogous to discovering the linked genes from chromosomes or related information. Therefore, by regarding each linkage between genes as an association rule, this study adopts data mining to learn these rules. The rest of this paper is organized as follows. Section 2 introduces the proposed linkage mining approach. Section 3 summarizes the experimental results. Conclusions are finally drawn in Section 4, along with recommendations for future research.

2. Linkage Mining

As an implementation of data mining for linkage discovery, this study adopts the well-known Apriori algorithm to mine the data generated by D^5 for association rules. The set of items in an association rule corresponds to the set of loci in a building block. The disruptiveness caused by crossover can be resolved using information on building blocks. The major components for mining linkage are described below.

2.1 D^5 Algorithm

The D^5 algorithm [13] integrates the advantages of PMs and EDAs. For each locus, D^5 takes three steps:

Digital Object Identifier 10.1109/MCI.2009.935310

1556-603X/10/$26.00©2010IEEE

1) *Perturb* genes at that locus for fitness difference.
2) *Cluster* the population according to fitness difference.
3) *Estimate* the linkage from each cluster.

In the first step, D^5 perturbs (i.e. $0 \rightarrow 1$ or $1 \rightarrow 0$) all genes at locus i in the population and calculates the resulting difference in fitness, that is,

$$\Delta f_i(\mathbf{c}) = f(\mathbf{c}') - f(\mathbf{c}),$$

where $f(\mathbf{c})$ and $f(\mathbf{c}')$ denote the fitness values before and after perturbing the gene at locus i, respectively.

Based on fitness difference, D^5 partitions the population into several sub-populations C_{i1}, C_{i2}, ..., C_{iN_i}. For a sub-population C_{ij} larger than a predefined size, D^5 constructs the largest set of loci, denoted by V_{ij}, that contains locus i and holds the minimum entropy for the genes at these loci. The set V_{ij} serves as an estimated dependency set for further deriving linkage sets, i.e. building blocks. The derivation of linkage sets in D^5 depends on the following problem types: For non-overlapping functions, linkage set V_i for locus i is the estimated dependency set V_{ij^*} with the smallest entropy, i.e. $j^* = \text{argmin}_j E(V_{ij})$, where $E(\ )$ refers to the entropy. As for overlapping functions, the enhanced D^5 [16] investigates the mutual dependency relationship among the unions of estimated dependency sets $\widetilde{V}_i = \bigcup_j V_{ij}$ for all locus i to determine the linkage sets.

This study examines the feasibility of using the Apriori algorithm to mine the estimated dependency sets for building blocks. This mining approach can identify building blocks precisely and improve the performance of GAs on both the non-overlapping and overlapping functions.

2.2 Apriori Algorithm

Association rules are commonly expressed in the form of $X \rightarrow Y$, representing whenever itemset (i.e. set of items) X is in a transaction, itemset Y is also in the transaction. With respect to the metaphor in Fig. 1, the elements (loci) of an estimated dependency set are regarded as the items of a transaction. Moreover, association rules refer to the strong connections of certain items, which is analogous to the linkage relationship of particular genes. Based on this analogy, this study applies the Apriori algorithm, a well-established method for mining association rules, to mine the estimated dependency sets generated by D^5 for building blocks in GAs.

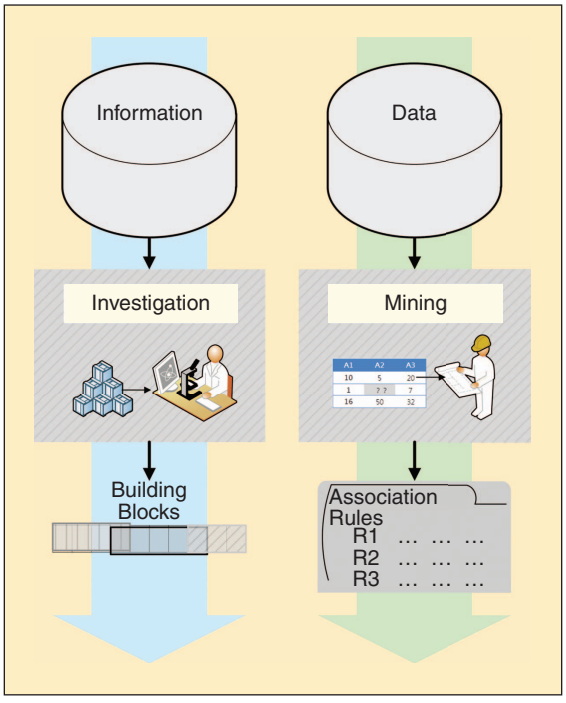


FIGURE 1 Metaphor connecting linkage discovery to data mining.

Some related terms in data mining are defined before introducing the Apriori algorithm. The *support* of an itemset X represents the number of transactions in which the members of X all exist. Additionally, C_k denotes the collection of candidate itemsets of size k, and L_k represents the collection of large itemsets of size k, which have a support exceeding the predefined minimum support *minsup*.

The Apriori algorithm was proposed by Agrawal et al. [17, 18]. Starting with $k = 1$, Apriori incrementally generates candidate itemsets C_k and finds out large itemsets L_k from C_k, until L_k is empty. This algorithm can effectively reduce the computational effort in checking various combinations of itemsets for large itemsets.

Figure 2 illustrates use of the Apriori algorithm to mine for linkage. The database consists of four transactions,

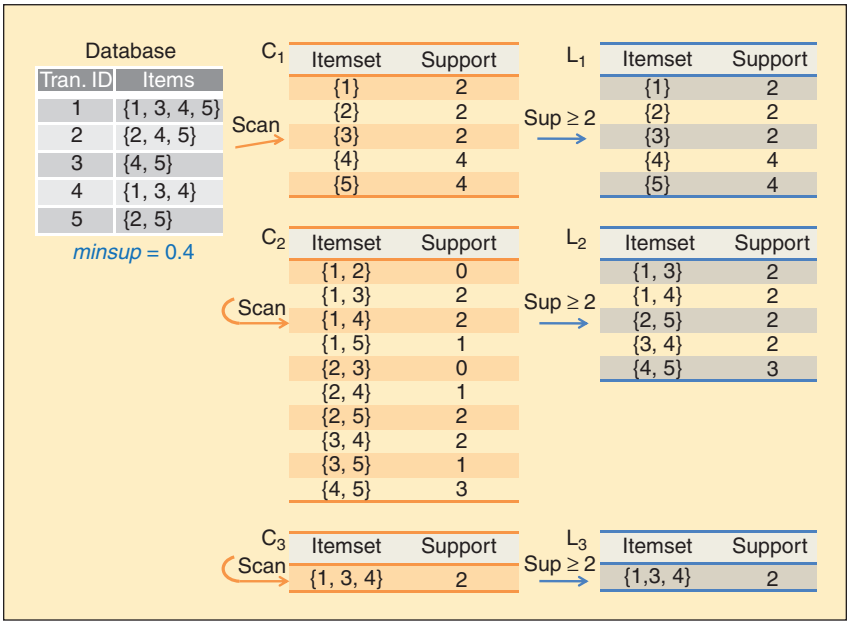


FIGURE 2 Generation of candidate itemsets and large itemsets.

Instead of applying GAs for data mining, this study addresses linkage discovery, an essential topic in GAs, by using data mining methods.

where the first transaction represents an estimated dependency set, including loci $\{1, 3, 4, 5\}$. Apriori first scans for candidate itemsets C_1 and then determines large itemsets L_1 from them. In the example, given (number of transactions) $\times$ *minsup* $= 2$, only the itemsets in C_1 that appear at least twice among all transactions in the database are placed into L_1. The candidate itemsets C_2 are formed by joining and pruning the itemsets in L_1. Apriori continues with this procedure to generate C_2, L_2, C_3, and L_3. We posit that each large itemset is a building block, except for the subsets of higher-order itemsets. Hence, the large itemsets $\{2, 5\}$, $\{4, 5\}$, and $\{1, 3, 4\}$ are obtained as building blocks.

TABLE 1 Success rate (SR) and the number of evaluations (NES) to achieve the optimal solutions for D^5 and the proposed mining approach on the $trap_5$ function with chromosome length *l* and population size *p*. Boldface signifies a superior SR or a significantly lower NES.

		SR (%)		NES	
l	*p*	D^5	MINING	D^5	MINING
200	400	50.0	**100.0**	132,840	**108,560**
400	400	10.0	**80.0**	235,880	**218,360**
600	600	60.0	**100.0**	447,240	451,620
800	800	50.0	**100.0**	770,400	**742,800**
1,000	1,000	70.0	**100.0**	1,157,500	**1,140,000**

3. Experimental Results

This study conducts experiments on the non-overlapping $trap_5$ function and the function with stochastic circularly overlapping sub-functions (SCOS) [16]. Performance of the proposed mining approach based on Apriori is compared with that of D^5. The GA uses the context dependent crossover (CDC) [16] and no mutation for both methods. Notably, the population size is the same for both the linkage discovery phase and evolution phase. The minimum support is empirically set in the range of $[0, 0.1]$. Each experimental setting includes 30 independent runs of 200 generations concerning the stochastic nature of GAs. Performance measures include success rate (SR) and the number of evaluations (NES) to achieve the optimal solutions for all 30 runs.

Table 1 compares the performances of D^5 and the proposed mining approach on the trap function. The mining approach yields 100% SR on all test instances except for $l = 400$ and outperforms D^5 in SR on all test instances. Additionally, the approach requires significantly fewer evaluations to achieve the optima solutions than D^5 does, where the statistical significance is examined by a one-tailed *t*-test with a confidence level of 0.05. The preferable experimental results demonstrate the advantage of the proposed mining approach over D^5 in terms of solution quality and efficiency.

Furthermore, Table 2 presents the SR and NES for D^5 and the proposed mining approach on the SCOS function. Notably, a large variance σ^2 incurs highly complex and overlapping interactions between building blocks in the SCOS function. Experimental results indicate that the mining approach can achieve better SR with a lower NES than D^5 on most test instances. Moreover, the mining approach using a relatively small population is comparable with D^5 using a relatively large population in terms of SR on several instances. Regarding algorithmic efficiency, the mining approach uses a significantly lower NES than D^5 does. In addition to validating the capability of superior SR, these outcomes indicate that the mining approach can substantially decrease the required computational resources of D^5 with respect to population size and number of evaluations.

TABLE 2 Success rate (SR) and the number of evaluations (NES) to achieve the optimal solutions for D^5 and the proposed mining approach on the SCOS function with variance σ^2, chromosome length *l*, and population size *p*. Boldface signifies a superior SR or a significantly lower NES.

			SR (%)		NES	
σ^2	*l*	*p*	D^5	MINING	D^5	MINING
1^2	60	500	100.0	100.0	51,617	**46,867**
		1,000	100.0	100.0	84,000	83,400
	90	500	76.7	**93.3**	93,483	**81,950**
		1,000	100.0	100.0	122,400	122,200
	120	500	63.3	**83.3**	126,017	**112,333**
		1,000	100.0	100.0	161,700	161,833
2^2	60	500	13.3	**90.0**	125,317	**68,250**
		1,000	100.0	100.0	108,600	**88,233**
	90	500	0.0	**60.0**	145,500	**113167**
		1,000	70.0	**96.7**	204,133	**145,267**
	120	500	0.0	**23.3**	160,500	**148,650**
		1,000	46.7	**93.3**	276,800	**183,967**
5^2	60	1,000	46.7	**96.7**	216,633	**122,367**
		2,000	100.0	100.0	213,467	**199,000**
	90	1,000	0.0	**50.0**	291,000	**249,267**
		2,000	86.7	**90.0**	386,933	360,667
	120	2,000	16.7	**60.0**	622,267	**550,800**
		3,000	63.3	**86.7**	801,400	**741,600**
10^2	60	1,000	46.7	**100.0**	211,933	**104,467**
		2,000	100.0	100.0	199,133	195,533
	90	1,000	0.0	**53.3**	291,000	**238,500**
		2,000	70.0	**83.3**	428,000	**374,200**
	120	2,000	23.3	**60.0**	619,133	**534,867**
		3,000	70.0	**83.3**	817,600	**756,000**

Learning association rules involves finding the implicitly associated items from transactions, which is analogous to discovering the linked genes from chromosomes or related information.

4. Conclusions

Instead of applying GAs to data mining, this study identifies the linkage relationship in GAs through data mining. Based on the analogy between building blocks and association rules, this study utilizes the Apriori algorithm to mine the estimated dependency sets generated by D^5 for the large itemsets as building blocks.

Experimental results of non-overlapping and complex overlapping functions indicate that the proposed mining approach can improve D^5 in terms of solution quality and efficiency. Specifically, the mining approach achieves a higher success rate with significantly fewer evaluations to achieve the optimal solutions than D^5 does. With a smaller population, the mining approach can also yield a success rate comparable with D^5 using a larger population. These superior outcomes validate the effectiveness and efficiency of the proposed approach in linkage discovery.

In addition to enhancing D^5, linkage mining paves the way for a new field of research. We recommend the following directions for future research. First, future studies should attempt to obtain the data to be mined for building blocks from methods other than D^5. Second, given the availability of a considerable number of association rules learning methods, more effective methods should be adopted for linkage mining applications. Finally, as well as association rules, other data mining tasks, such as classification and clustering, are highly promising for use in linkage discovery.

Acknowledgments

The authors would like to thank Miwako Tsuji for the source code of D^5 and CDC.

References

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Mateo, CA: Morgan Kaufmann, 2000.

[2] A. A. Freitas, *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Berlin: Springer-Verlag, 2002.

[3] S. Mitra, S. K. Pal, and P. Mitra, "Data mining in soft computing framework: A survey," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 3–14, 2002.

[4] J. Holland, *Adaptation in Natural and Artificial Systems*. Univ. of Michigan Press, 1975.

[5] D. E. Goldberg, *The Design of Innovation: Lessons From and For Competent Genetic Algorithms*. Berlin: Springer-Verlag, 2002.

[6] M. Pelikan, D. E. Goldberg, and E. Cantú-Paz, "BOA: The Bayesian optimization algorithm," in *Proc. 1999 Genetic and Evolutionary Computation Conf.*, 1999, pp. 525–532.

[7] H. Mühlenbein and T. Mahnig, "FDA – A scalable evolutionary algorithm for the optimization for the optimization of additively decomposed functions," *Evolution. Computat.*, vol. 7, no. 4, pp. 353–376, 1999.

[8] G. R. Harik, "Linkage learning via probabilistic modeling in the ECGA," *Univ. of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, Tech. Rep.*, 1999.

[9] M. Munetomo and D. E. Goldberg, "Designing a genetic algorithm using the linkage identification by nonlinearity check," *Univ. of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, Tech. Rep.*, 1998.

[10] M. Munetomo and D. E. Goldberg, "Identifying linkage by nonlinearity check," *Univ. of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, Tech. Rep.*, 1998.

[11] M. Munetomo and D. E. Goldberg, "Linkage identification by non-monotonicity detection for overlapping functions," *Evolution. Computat.*, vol. 7, no. 4, pp. 377–398, 1999.

[12] M. Munetomo, "Linkage identification based on epistasis measures to realize efficient genetic algorithms," in *Proc. 2002 IEEE Congress on Evolutionary Computation*, 2002, pp. 1332–1337.

[13] M. Tsuji, M. Munetomo, and K. Akama, "Modeling dependencies of loci with string classification according to fitness differences," in *Proc. 2004 Genetic and Evolutionary Computation Conf.*, 2004, pp. 246–257.

[14] M. Tsuji and M. Munetomo, "Linkage analysis in genetic algorithms," in *Computational Intelligence Paradigms*. Berlin: Springer-Verlag, 2008, pp. 251–279.

[15] Y. P. Chen, T. L. Yu, K. Sastry, and D. E. Goldberg, "A survey of genetic linkage learning techniques," *Univ. of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, Tech. Rep.*, 2007.

[16] M. Tsuji, M. Munetomo, and K. Akama, "A crossover for complex building blocks overlapping," in *Proc. 2006 Genetic and Evolutionary Computation Conf.*, 2006, pp. 1337–1344.

[17] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. 1993 ACM SIGMOD Int. Conf. Management of Data*, 1993, vol. 22, no. 2, pp. 207–216.

[18] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. Very Large Data Bases*, 1994, vol. 1215, pp. 487–499.

Society Briefs *(continued from page 4)*

using the appropriate IEEE spreadsheet available on the link above (budget forms are not required for technical cosponsorship proposals). Proposals should be submitted 18 months in advance of the desired conference dates, especially with regards to conferences with requested financial support from IEEE CIS. Both the IEEE CIS conferences committee and the IEEE CIS administrative committee must approve all budgets (and budget revisions).

AsVP Conferences, it is my goal to encourage continued growth in our IEEE CIS sponsored conferences. This can be accomplished through new special sessions at standard conference events, or by turning very popular special sessions into their own symposia. Fostering additional interaction between the IEEE CIS technical committees and conferences will help increase quality and attendance. I also feel that it is very important that students be attracted to these events and look forward to IEEE CIS continuing its assistance with student travel grants. If you have any thoughts or questions about the conference process please contact me at gfogel@natural-selection.com.