Gene association analysis for Bayesian network integrated classification

Sha Liao, Xuhong Liao*, Zhijie Li
School of Information Science and Engineering, Hunan Institute of Science and Technology, Yueyang 414006, Hunan, China

ABSTRACT

This paper proposes a gene association analysis algorithm that effectively identifies causal relationships between genes through gene association entropy, and uses heuristic search strategies to construct gene association Bayesian tree (GABT) and gene association Bayesian forest (GABF). Unlike ordinary gene Bayesian networks that describe the dependency relationship between gene expression levels, GABT and GABF are a type of gene sequence Bayesian network. The object of gene association analysis is the sequence formed by sorting the gene expression values of biological tissue samples and replacing them with gene column subscripts. The experimental results on multiple tumor or non tumor gene expression datasets show that the Bayesian network classification algorithm based on gene association analysis can better fit gene expression data than other similar algorithms, with significantly improved accuracy or reduced analysis time.

Keywords: Gene expression data, gene association entropy, bayesian network, gene association bayesian tree, integrated classification

1. INTRODUCTION

Gene expression data is a specific type of big data with biological background. Gene expression data analysis covers areas such as unsupervised learning, supervised learning, and gene regulatory networks, among which gene expression data classification is the most important supervised learning method[1,2]. Due to the special subspace pattern similarity of gene expression data, in order to mine pattern information, the gene expression values of tissue samples are often sorted and replaced with column labels. The order preserving submatrix (OPSM)[3,4] is a typical biclustering method for mining its longest common subsequence.

After sorting gene expression values and replacing them with gene labels, the gene sequence forms a hidden Markov model, and the state transition probability of this model implies causal relationships between genes[5,6]. Mining causal structures hidden in training data and applying them to Bayesian networks is popular research directions in recent years.

Bayesian networks[7] typically use the network topology of naive Bayes (NB) models as the basic framework for classifiers. Due to the absence of conditional independence assumptions and any causal relationships in NB, there is no need for network structure learning, which often does not match the actual situation. Therefore, a restricted Bayesian network classifier adds directed edges between nodes to improve NB. Restricted Bayesian network classifiers[8-10] mainly include single and ensemble Bayesian network classifiers. Single structure Bayesian network classifiers include KDB, TAN, CFWNB, BCT, etc. Bayesian network ensemble classifiers include AODE, WATAN, IWAODE, WAODE-MI, TAODE, BCF, etc.

However, existing Bayesian network classifiers or ensemble classifiers cannot be directly used for gene expression data classification[11-13]. (1) The use of Euclidean distance does not take into account the special similarity measurement criteria for gene expression data. The expression values of closely related genes in biology may not be close, but they may exhibit a consistent trend of rising and falling at the same time; (2) The existing Bayesian network classifiers have high time complexity and poor performance when the number of variables increases, resulting in time bottlenecks; (3) At present, Bayesian network classifiers are aimed at general discrete feature variables, and their performance is poor when directly classifying gene expression data.

*lxh_2402@163.com

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 133953C · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3049912

Based on the above issues, it is urgent to propose an advanced Bayesian network method to handle the problem of gene expression data classification. This paper studies gene association analysis algorithms and uses heuristic search strategies to construct a gene association Bayesian forest classifier. The experimental results verify the effectiveness of the algorithm proposed in this paper.

2. MINING GENE ATOMIC SEQUENCES OF GENE EXPRESSION DATA

With the development of genomics and bioinformatics, a massive amount of gene expression data related to various diseases has been accumulated[14]. Table 1 is an example of gene expression data for classification. Among them, each row represents a tissue sample s_i (s_{i1}, s_{i1}, …, s_{in}). Each row in the table can be regarded as a vector, where s_{ij} is the expression level of gene j in sample s_i. If Table 1 is a classified dataset, the tissue sample can be represented as s_i (s_{i1}, s_{i1}, …, s_{in}, y_i), where y_i is the category label to which the sample belongs, such as "-", "+", etc.

Table 1. Sample example of gene expression data.

Sample	G_1	G_2	G_3	G_4	G_5	G_6
s_1(-)	0.155	0.076	-0.201	0.254	0.013	-0.181
s_2(-)	0.217	0.084	0.150	0.165	-0.159	0.132
s_3(-)	0.375	0.115	0.284	0.076	-0.094	0.155
s_4(-)	0.238	0	-0.159	0.129	-0.191	0.217
s_5(-)	-0.073	-0.146	0.443	0.818	-0.341	0.227
s_6(-)	0.394	0.909	0.426	0.768	1.070	0.226
s_7(+)	0.385	0.822	0.244	0.550	1.013	0.327
s_8(+)	0.329	0.690	0.066	0.529	0.790	0.313
s_9(+)	0.384	0.730	0.066	0.529	0.852	0.313
s_{10}(+)	-0.316	-0.191	0.202	-0.140	0.043	0.076

2.1 Mining frequent gene atomic sequences

In order to explore the gene correlation of the classified gene expression data shown in Table 1, we first ignore the sample categories and preprocess the gene expression values by sorting them. This transforms pattern mining into a frequent order-preserving sequence mining problem[15]. Here, we mainly consider the frequent atomic sequence mining problem with a length of 2.

(1) Sorting the gene expression values of each sample in descending order, as shown in Table 2.

(2) Replacing gene expression values with gene column subscripts, as shown in Table 3.

(3) Counting the number of occurrences of frequent gene atomic sequences.

If the minimum support number is 2, the statistics of frequent gene atomic sequences and their occurrence times are shown in Table 4.

Table 2. Descending sorting of gene expression values in Table 1.

Sample	Descending sorting of gene expression values					
s1(-)	0.284(g3)	0.155(g1)	0.097(g4)	0.076(g2)	0.023(g6)	0.013(g5)
s2(-)	0.409(g3)	0.217(g1)	0.138(g4)	0.129(g6)	0.084(g2)	-0.159(g5)

Sample	Descending sorting of gene expression values					
s3(-)	0.375(g1)	0.254(g4)	0.115(g2)	-0.094(g5)	-0.181(g6)	-0.201(g3)
s4(-)	0.238(g1)	0.165(g4)	0.15(g3)	0.132(g6)	0.0(g2)	-0.191(g5)
s5(-)	0.442(g3)	0.063(g6)	-0.073(g1)	-0.077(g4)	-0.146(g2)	-0.341(g5)
s6(-)	1.070(g5)	0.909(g2)	0.818(g4)	0.443(g3)	0.394(g1)	0.227(g6)
s7(+)	1.013(g5)	0.822(g2)	0.768(g4)	0.426(g1)	0.385(g6)	0.226(g3)
s8(+)	0.790(g5)	0.690(g2)	0.55(g4)	0.329(g1)	0.327(g6)	0.244(g3)
s9(+)	0.852(g5)	0.730(g2)	0.529(g4)	0.384(g1)	0.313(g6)	0.066(g3)
s10(+)	0.202(g3)	0.076(g6)	0.043(g5)	-0.140(g4)	-0.191(g2)	-0.316(g1)

Table 3. Gene column subscript list.

Sample	Gene column subscript sequence
s1(-)	3→1→4→2→6→5
s2(-)	3→1→4→6→2→5
s3(-)	1→4→2→5→6→3
s4(-)	1→4→3→6→2→5
s5(-)	3→6→1→4→2→5
s6(-)	5→2→4→3→1→6
s7(+)	5→2→4→1→6→3
s8(+)	5→2→4→1→6→3
s9(+)	5→2→4→1→6→3
s10(+)	3→6→5→4→2→1

Table 4. Frequent gene atomic sequences.

Atomic	Counts	Atomic	Counts
6→5	2	4→1	2
4→3	3	3→1	4
2→4	4	4→2	4
1→4	5		
2→5	4		
3→6	3		
1→6	4		
6→2	2		
6→3	3		
5→2	4		

3. GENE ASSOCIATION ANALYSIS

Gene association analysis draws inspiration from the ideas of frequent patterns and association rules in data mining, and defines gene association rules for mining implicit causal relationships using gene association entropy.

3.1 Defining gene association entropy

Defination 1. Gene association entropy. For any frequent gene atomic sequence $x \to y$, we let $Y_i(i=1,2,\ldots,n)$ be the y parent node gene, and $X_j\,(j=1,2,\ldots,m)$ be the x parent node gene. The correlation entropy

$$H(x \to y) = \sum_{X_j}\sum_{Y_i} H\left(Y_i \to y \middle| X_j \to x\right) = -\sum_{X_j}\sum_{Y_i} P\left(X_j \to x, Y_i \to y\right) \ln P\left(Y_i \to y \middle| X_j \to x\right)$$

Where, the calculation formula for conditional probability $P\left(Y_i \to y \middle| X_j \to x\right)$ is:

$$P\left(Y_i \to y \middle| X_j \to x\right) = \frac{P\left(X_j \to x, Y_i \to y\right)}{P\left(X_j \to x\right)} \tag{1}$$

In order to avoid the numerator or denominator of equation (1) being 0, the initial values of counters $c\left(X_j \to x, Y_i \to y\right)$ and $c\left(X_j \to x\right)$ in Table 3 are set to 1. The correlation entropy results of frequent gene atomic sequences are shown in Table 5.

Table 5. Association entropy of frequent gene atomic sequences.

No.	Atomic sequence	Correlation entropy	No.	Atomic sequence	Correlation entropy
1	5→2	0.805	8	4→2	1.257
2	4→3	0.852	9	2→4	1.318
3	2→5	0.856	10	3→6	1.386
4	6→5	0.946	11	6→2	1.453
5	6→3	1.007	12	4→1	1.568
6	3→1	1.109	13	1→6	2.047
7	1→4	1.159			

3.2 Genetic association rule mining algorithm

The gene association rule mining algorithm measures the correlation degree of frequent gene atomic sequences through association entropy, and then sorts and compares the gene association entropy $H(x \to y)$ to obtain a set of gene association rules $x \to y$ with strong correlation degree. The pseudocode of the gene association rule mining algorithm (GA) is shown in Algorithm 1.

Algorithm 1. ***GA*** ($\boldsymbol{A}$, σ , *max_entropy*)

Input: gene expression data-$\boldsymbol{A}$, minimum support-σ , correlation entropy threshold-*max_entropy*

Output: strong gene association rule set-G_{rule}

(1) G_{rule} = null

(2) $\mathcal{A}$ =ordering($\boldsymbol{A}$, G) // G is the set of gene column labels

(3) **for each** S∈$\mathcal{A}$ **do**

(4) **for each** $x \to y$∈S **do**

(5) **if** $x \to y \notin G_{rule}$ **then**

(6) $w(x \to y)=1$, G_{rule}.add($x \to y$)

(7) **else** $w(x \rightarrow y) = w(x \rightarrow y) + 1$

(8) **for each** $x \rightarrow y \in G_{rule}$ **do**

(9) **if** $c(x \rightarrow y) \geq \sigma$ **then**

(10) Calculating $H(x \rightarrow y)$ according to Definition 1

(11) **else** G_{rule}.delete($x \rightarrow y$)

(12) ordering(G_{rule}, $H(x \rightarrow y)$)

(13) G_{rule}= G_{rule}.intercept(*max_entropy*)

(14) **return** G_{rule}

4. INTEGRATED CLASSIFICATION OF GENE BAYESIAN ASSOCIATION NETWORKS

4.1 Gene Bayesian association tree

Bayesian networks based on gene association analysis are an effective method for constructing network models. Due to the asymmetric nature of gene association ($g_i \rightarrow g_j$), it can be well used to analyze the causal relationship between genes g_i and g_j[16]. In order to control the complexity of the model, this paper limits the topology of the gene Bayesian network to a first-order correlated directed acyclic graph, where any gene g_i has only one parent gene F_i, forming a gene Bayesian association tree (GBAT).

Algorithm 2 is a pseudocode for constructing a gene association tree algorithm, which describes the process of adding directed edges to GBAT based on gene association degree.

Algorithm 2 GBAT_Learning (G_{rule}, $\mathbf{g}=\{g_1, g_2, \ldots, g_n, C\}$)

Input: Gene Association Rule Set -G_{rule}，Genes and Categories -$\{g_1, g_2, \ldots, g_n, C\}$

Output: GBAT network topology

(1) Initializing GBAT tree: T(r)=(U, V)，U={C}, V=null

(2) Root selection: $U=U \cup \{g_r\}$, $\mathbf{g}= \mathbf{g} \backslash \{g_r\}$, $V = V \cup \{C \rightarrow g_r\}$

(3) **while** ($\mathbf{g} \neq \varnothing$)

(4) Selecting the maximum correlation $g_i \rightarrow g_j (g_i \in U, g_j \in \mathbf{g})$

(5) $U=U \cup \{g_j\}$，$\mathbf{g}= \mathbf{g} \backslash \{g_j\}$，$V = V \cup \{C \rightarrow g_j, g_i \rightarrow g_j\}$

(6) **return** T(r)

Defination 2. Gene conditional probability table (GCPT). For any gene node Y in the Bayesian correlation tree, if $X_i (i=1,2,\ldots,n)$ is its parent node gene, then the value of gene node Y is $\{X_i \rightarrow Y, c \rightarrow Y\}$. If Z_i is the parent node of X_i, then in the conditional probability table of gene node Y, the conditional probabilities for row $<c, Z_i \rightarrow X_i>$, and column $P(Y=X_i \rightarrow Y)$ are represented as

$$P(Y=X_i \rightarrow Y | c, Z_i \rightarrow X_i) \quad (2)$$

The frequency of association rules is used to represent the corresponding conditional probability of GCPT, where c represents the class label.

4.2 Genetic Bayesian association forest

If different genes are selected as root nodes for gene association inference, there will be significant differences in the structure of GBAT trees constructed from training data, reflecting the diversity of gene association relationships between different GBAT trees. The diversity of GBAT increases the generalization classification ability of the ensemble model. This paper further constructs a gene Bayesian association forest (GBAF) classifier, as shown in Figure 1.

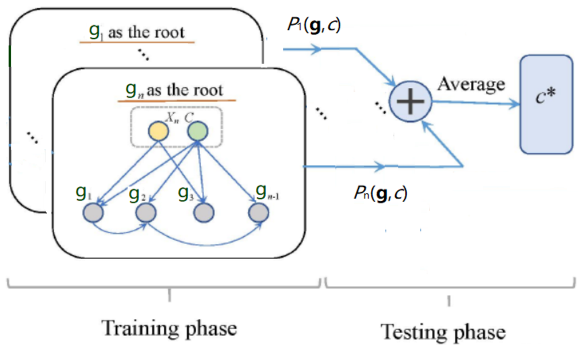

Figure 1. The learning framework of GBAF.

Algorithm 3 is a pseudo code for the training and testing process of GBAF classifiers.

Algorithm 3 GBAF_ Learning($\mathcal{D}$, $\mathbf{g}$={g_1, g_2, …, g_n, C}, $\boldsymbol{x}$)

Input: Gene expression data - $\mathcal{D}$, Genes and categories-{g_1, g_2, …, g_n, C}, Test sample-$\boldsymbol{x}$

Output: Predicted class labels -y^*

GBAF_ Training($\mathcal{D}$, $\mathbf{g}$={g_1, g_2, …, g_n, C})

(1) G_{rule} = $\boldsymbol{GA}$ ($\mathcal{D}$, σ , *max_entropy*)

(2) **for** (r =1 to n)

(3) selecting g_r as root node

(4) GBAT_ Learning(G_{rule}, $\mathbf{g}$={g_1, g_2, …, g_n, C})

(5) **return** $GBAT_1$, …, $GBAT_n$

GBAF_ Testing($GBAT_1$, …, $GBAT_n$, $\boldsymbol{x}$)

(6) **for** (k =1 to m)

(7) **for** (r =1 to n)

(8) computing $P_r(\boldsymbol{x}, c_k)$ of $GBAT_r$ according to Figure 1.

(9) $P(\boldsymbol{x}, c_k) = \frac{1}{n}\sum_{r=1}^{n} P_r\left(\boldsymbol{x}, c_k\right)$

(10) **return** $y^* = \arg\max_{c_k \in C} \frac{P\left(\boldsymbol{x}, c_k\right)}{\sum_{k=1}^{m} P\left(\boldsymbol{x}, c_k\right)}$

---

5. EXPERIMENTAL RESULTS AND ANALYSIS

This paper evaluates the performance of the proposed algorithm using 9 datasets shown in Table 6. The experiment was conducted on a computer with a 2.60GHz Intel (R) Core (TM) i7-6700HQ CPU, 16GB of memory, and Windows 10 operating system.

5.1 Datasets and comparison classifiers

The 9 gene expression data used in experiments include 6 tumor datasets and 3 non tumor datasets, mainly from libSVM (http://www.csie.ntu.edu.tw/~Cjlin/libsvmtools/datasets/) and UCI website (https://archive.ics.uci.edu/ml/datasets). Tumor data sets include Leukemia, Colon, SRBCT, Brain, Breast cancer and Duke_bc; Non tumor datasets include Heart, Mushrooms, and Proteins. Table 6 lists the parameters of the relevant dataset.

Table 6. Gene expression dataset.

Dataset	No. of genes	No. of samples	No. of class
Leukemia	7129	72	2
Colon	2000	62	2
SRBCT	2308	83	4
Brain	5920	90	5
Breast cancer	10	683	2
Duke_bc	7129	44	2
Heart	13	270	2
Mushrooms	112	8124	2
Protein	357	17766	3

The experiment compared the gene Bayesian association forest classification algorithm GBAF proposed in this paper with Bayesian network variants and other classifier algorithms, and verified the ensemble effectiveness of GBAF.

- Variations of Bayesian networks

Naive Bayesian Network BN_NB, two conditional independence testing algorithms BN_CI and BN_ICS, simulated annealing global scoring metric BN_SA, CFWNB[17] (correlation based feature weighting filter for naïve Bayes), WATAN[18] (weighted average tree augmented naïve Bayes), and the GBAF proposed in this paper.

- Other classifier algorithms

SVM(Support Vector Machine), KNN(K-Nearest Neighbor), LR(Logistic Regression), LB(LevBag), OB(OzaBoost), RF(Random Forests)

5.2 Comparison results of BN variant algorithms

(1) RMSE experimental results

Table 7 presents the experimental results of the RMSE metric for BN variant classifiers on 9 datasets.

Table 7. RMSE experiment results.

Dataset	BN_NB	BN_SA	CFWNB	WATAN	BN_CI	BN_ICS	GBAF
Leukemia	0.4830	0.5045	0.4111	0.4277	0.3952	0.4056	0.3696
Colon	0.4025	0.3716	0.2952	0.3315	0.3237	0.3409	0.3174
SRBCT	0.0689	0.0137	0.0270	0.0177	0.0159	0.0124	0.0331
Brain	0.3020	0.2759	0.2419	0.2705	0.2304	0.2501	0.2341
Breast cancer	0.2613	0.2800	0.3589	0.3203	0.3201	0.3198	0.2491
Duke_bc	0.4915	0.4526	0.3150	0.3076	0.3250	0.3297	0.3078
Avg RMSE	0.3349	0.3164	0.2749	0.2792	0.2684	0.2764	0.2652
Avg rank	5.833	4.833	3.833	4.000	2.833	3.500	3.167
Heart	0.6005	0.4791	0.3384	0.3418	0.3450	0.3443	0.3285
Mushrooms	0.3495	0.2315	0.4334	0.4023	0.3992	0.3984	0.3161

Dataset	BN_NB	BN_SA	CFWNB	WATAN	BN_CI	BN_ICS	GBAF
Protein	0.4671	0.2892	0.3929	0.3504	0.3516	0.3487	0.3397
Avg RMSE	0.4724	0.3333	0.3882	0.3648	0.3653	0.3638	0.3548
Avg rank	6.333	3.567	5.000	4.333	5.000	3.667	2.000
Overall RMSE	0.3807	0.3220	0.3126	0.3078	0.3007	0.3055	0.2950
Overall rank	6.167	5.111	4.222	4.111	3.556	3.556	2.778

(2) Friedman and Nemenyi test

The seven algorithms used in the experiment follow F distribution with degrees of freedom of 7-1=6 and (7-1) × (9-1)=48. When a =0.05, the critical value of $F(6, 48)$ is 2.298. From Table 7, it can be calculated that F_F is 3.337, and the result is greater than 2.298. Therefore, Nemenyi's subsequent test is conducted.

When a=0.05, the critical value CD of 7 algorithms on 9 datasets is 2.408. As shown in Figure 2, GBAF outperforms other algorithms in RMSE.

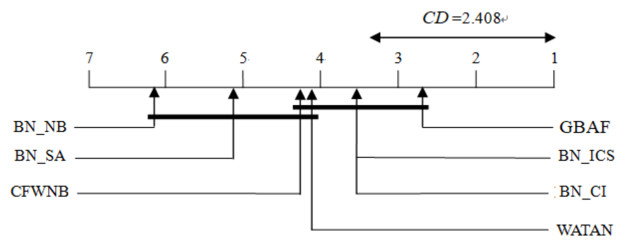

Figure 2. Comparison results of Nemenyi test in RMSE.

(3) Classification time comparison

The comparison results of classification time in Figure 3 show that the training time of GBAF is slightly longer than CFWNB and BN-NB, and the testing time is only slightly longer than CFWNB. Therefore, GBAF has good time performance.

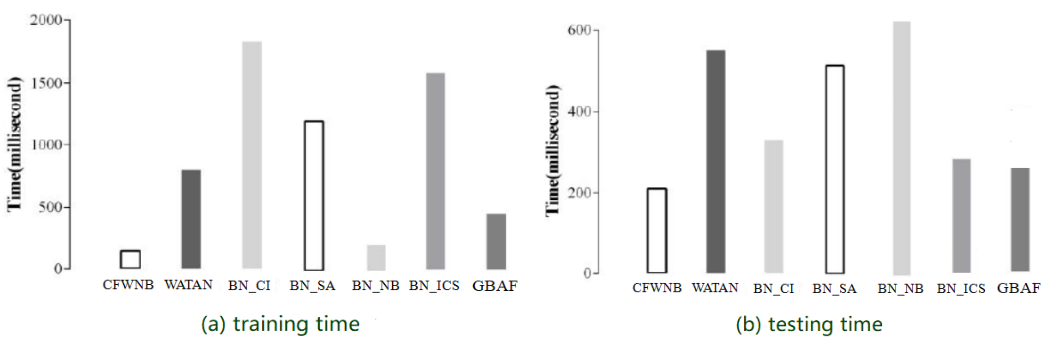

Figure 3. Comparison of average training and classification time on 9 datasets.

5.3 Comparison between GBAF and other classifiers

We compare the accuracy metrics of GBAF with the other six classifiers and provide the Nemenyi test statistical analysis results.

(1) Accuracy experimental results

Table 8 presents the experimental results of classification accuracy indicators for GBAF and 6 other classifiers on 9 datasets.

Table 8. Accuracy experimental results.

Dataset	LR	SVM	KNN	OB	LB	RF	GBAF
Leukemia	0.7667	0.7444	0.8019	0.7537	0.8074	0.8120	0.8130
Colon	0.7570	0.7897	0.8272	0.7804	0.7804	0.7477	0.8037
SRBCT	0.9834	0.9993	0.9980	0.9986	0.9989	0.9992	0.9992
Brain	0.8176	0.8109	0.8514	0.8311	0.8581	0.8446	0.8581
Breast cancer	0.8976	0.8656	0.7504	0.7264	0.7168	0.7168	0.7488
Duke_bc	0.7584	0.7951	0.8501	0.8620	0.8498	0.8442	0.8637
Avg Acc	0.8301	0.8342	0.8465	0.8254	0.8352	0.8274	0.8478
Avg rank	6.333	5.333	3.333	4.583	3.917	4.500	2.000
Heart	0.6395	0.7687	0.8402	0.8265	0.8401	0.8401	0.8435
Mushrooms	0.8779	0.9875	0.6900	0.7735	0.7338	0.7370	0.7714
Protein	0.6727	0.6660	0.6300	0.7248	0.7211	0.7257	0.7347
Avg Acc	0.7300	0.8074	0.7201	0.7749	0.7650	0.7676	0.7832
Avg rank	6.667	6.333	4.333	3.667	4.500	3.500	2.000
Overall Acc	0.7968	0.8252	0.8044	0.8086	0.8118	0.8075	0.8262
Overall rank	6.801	6.234	4.673	4.812	4.318	4.000	2.000

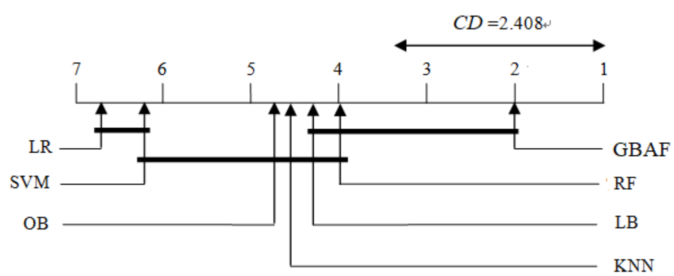

Figure 4. Comparison results of Nemenyi test in accuracy.

(2) Statistic analysis

The seven algorithms used in the experiment follow F distribution with degrees of freedom of 7-1=6 and (7-1)×(9-1)=48. When a =0.05, the critical value of F(6, 48) is 2.298. From Table 7, it can be calculated that F_F is 3.813, and the result is greater than 2.298. Therefore, Nemenyi's subsequent test is conducted.

When a=0.05, the critical value CD of 7 algorithms on 9 datasets is 2.408. As shown in Figure 4, GBAF outperforms other algorithms in accuracy.

6. CONCLUSION

For complex gene similarity relationships, traditional methods such as distance and correlation coefficient can only reflect gene linear similarity, while entropy based measurement methods such as conditional entropy and mutual information can mine pattern similarity and effectively reflect complex relationships between genes.

The gene association analysis algorithm proposed in this paper effectively identifies causal relationships between genes by defining gene association entropy, and uses heuristic search strategies to construct gene association Bayesian tree

GABT and gene association Bayesian forest GABF. How to apply gene association analysis to the construction of gene regulatory networks and the inference process of Bayesian networks is the focus of future research.

REFERENCES

[1] Yang, X. H., Wang, Z., Sun, J. and Xu, Z. B., "Unlabeled data driven cost-sensitive inverse projection sparse representation-based classification with 1/2 regularization," Science China: Information Sciences, 65(8), 1-18 (2022).

[2] Jiang, T. and Li, Z. H., "A survey on local pattern mining in gene expression data," Journal of Computer Research and Development, 55(11), 2343-2360 (2018).

[3] Wang, H. X., Pei, J. and Yu, P. S., "Pattern-based similarity search for microarray data," Proc. 11th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining, New York, USA, ACM, 814-819 (2005).

[4] Liu, J. Z. and Wang, W., "OP-clustering by tendency in high dimensional space," Proc. 3rd IEEE Int Conf on Data Mining, Piscataway, 187-194 (2003).

[5] Fan, H. Q., [Research of Bayesian causal forest based on ensemble learning], Changchun: Jilin University, Master's Thesis, (2022).

[6] Tsamardinos, I., Aliferis, C. F. and Statnikov, A., "Algorithms for large scale Markov Blanket discovery," Proc. 16th International Florida Artificial Intelligence Research Society Conference, AAAI Press, 376-381 (2003).

[7] Jiang, L., Zhang, L., Yu, L. and Wang, D., "Class-specific attribute weighted naïve Bayes," Pattern Recognition, 88(3), 321-330 (2019).

[8] Friedman, N., Geiger, D. and Goldszmidt, M., "Bayesian network classifiers," Machine Learning, 29(1), 131-163 (1997).

[9] Webb, G. I., Boughton, J. R. and Wang, Z., "Not so naïve Bayes: aggregating one-dependence estimators," Machine Learning, 58(5), 5-24 (2005).

[10] Wang, L., Qi, S., Liu, Y., Lou, H. and Zuo, X., "Bagging k-dependence Bayesian network classifiers," Intelligent Data Analysis, 25(1), 641-667 (2021).

[11] Liu, Y., Wang, L. and Mammadov, M., "Learning semi-lazy Bayesian network classifier under the c.i.i.d assumption," Knowledge-Based Systems, 208(6), 106-112 (2020).

[12] Heckerman, D., "A Bayesian approach to learning causal networks," Advances in Decision Analysis: From Foundations to Applications, 150(9), 285-295 (2013).

[13] Kong, H., Shi, L., Wang, L., Liu, Y., Mammadov, M. and Wang, G., "Averaged tree-augmented one-dependence estimators," Applied Intelligence, 7, 1-17 (2021).

[14] Yu, L. and Ren, S. J., "Prediction of cancerous pathogenic genes based on network and gene differential expression information," Scientia Sinica Vitae, 53(1), 94-108 (2023).

[15] Wang, C. Y., Zhang, J., Wang, X. P., et al., "Pathogenic gene prediction algorithm based on heterogeneous information fusion," Front Genet, 11(5), 123-131 (2020).

[16] Sun, J., Taylor, D. and Bollt, E. M., "Causal network inference by optimal causation entropy," SIAM Journal on Applied Dynamical Systems, 14(3), 73-106 (2015).

[17] Jiang, L., Zhang, L., Li. C. and Wu. J., "A correlation-based feature weighting filter for naïve baayes," IEEE Transactions on Knowledge and Data Engineering, 31(3), 201-213 (2018).

[18] Jiang. L., Cai. Z., Wang. D. and Zhang. H., "Improving tree augmented naïve bayes for class probability estimation," Knowledge-Based Systems, 26(10), 239-245 (2012).