http://dx.doi.org/10.14400/JDPM.2013.11.11.409

Learning Distribution Graphs Using a Neuro-Fuzzy Network for Naive Bayesian Classifier

Xue-Wei Tian*, Joon S. Lim**

Dept. of Computer Engineering, Gachon University*,**

퍼지신경망을 사용한 네이브 베이지안 분류기의 분산 그래프 학습

전설위*, 임준식**

가천대학교 전자계산학과*,**

Abstract Naive Bayesian classifiers are a powerful and well-known type of classifiers that can be easily induced from a dataset of sample cases. However, the strong conditional independence assumptions can sometimes lead to weak classification performance. Normally, naive Bayesian classifiers use Gaussian distributions to handle continuous attributes and to represent the likelihood of the features conditioned on the classes. The probability density of attributes, however, is not always well fitted by a Gaussian distribution. Another eminent type of classifier is the neuro-fuzzy classifier, which can learn fuzzy rules and fuzzy sets using supervised learning. Since there are specific structural similarities between a neuro-fuzzy classifier and a naive Bayesian classifier, the purpose of this study is to apply learning distribution graphs constructed by a neuro-fuzzy network to naive Bayesian classifiers. We compare the Gaussian distribution graphs with the fuzzy distribution graphs for the naive Bayesian classifier. We applied these two types of distribution graphs to classify leukemia and colon DNA microarray data sets. The results demonstrate that a naive Bayesian classifier with fuzzy distribution graphs is more reliable than that with Gaussian distribution graphs.

Key Words : naive Bayesian classifier, learning distribution, neuro-fuzzy classifier, Gaussian distribution, fuzzy membership function distribution

요 약 Naive Bayesian classifiers 네이브 베이지안 분류기는 샘플 데이터로부터 쉽게 구현될 수 있는 강력하고도 많이 사용되는 형식의 분류기다. 그러나 강한 조건부 독립성으로 인하여 효율이 저하되는 분류 결과를 초래한다. 일반적으로 네이브 베이지안 분류기는 연속성을 가진 특징 데이터의 우도를 처리하기 위해 가우시안 분산을 사용한다. 속성들의 확률밀도는 항상 가우시안 분산에 적합한 것만은 아니다. 또 다른 형식의 분류기는 지도학습을 통해 퍼지 규칙과 퍼지집합을 학습할 수 있는 퍼지신경망이다. 퍼지신경망과 네이브 베이지안 분류기간에는 구조적 유사성을 가지고 있기 때문에 퍼지신경망으로 학습된 분산 그래프를 네이브 베이지안 분류기에 적용하고자 하는 방안이 본 연구의 목적이다. 따라서 네이브 베이지안 분류기에 가우시안 분산 그래프를 사용한 결과와 퍼지 분산 그래프를 사용한 결과를 비교하였다. 이를 위해 leukemia와 colon의 DNA 마이크로어레이 데이터를 적용하여 분류하였다. 네이브 베이지안 분류기에 퍼지 분산 그래프를 사용한 결과 가우시안 분산 그래프를 사용한 결과보다 더 신뢰성이 있음을 보여주었다.

본 연구는 지식경제부 및 정보통신산업진흥원의 IT융합 고급인력과정의 지원을 받아 수행되었음 (NIPA-2013-H0401-13-1001)
본 연구는 2012년도 교육부의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행되었음 (No. NRF-2012R1A1A2044134)

Received 2 October 2013, Revised 22 October 2013
Accepted 20 November 2013
Corresponding Author: Joon S. Lim (Dept. of Computer Engineering)
Email: jslim@gachon.ac.kr
ISSN: 1738-1916

1. INTRODUCTION

Classification is a statistical operation in which certain objects are arranged in groups or classes according to their characteristics (sometimes known as attributes) to find a training set. There are many approaches to classification, including decision trees, neural networks, support vector machines, and Bayesian network. The Bayesian approach is the most commonly used when dealing with uncertainty because it is based on probability theory [1]. A Bayesian Network (BN) is a graphical model that encodes probabilistic relationships of variables. Its main distinguishing feature from classical statistical inference approaches is the use of subjective or personal beliefs (prior probabilities) in the analysis [2, 3]. These probabilistic approaches make strong assumptions about how the data is generated, and posit a probabilistic model that embodies these assumptions. Naive Bayesian network (NBN) classifiers are very robust regarding irrelevant attributes, and classification takes into account evidence from many attributes to make the final prediction [3, 4, 5]. Naive Bayesian has been adapted to handle continuous attributes primarily using Gaussian distributions or discretizing the domain, both of which present certain disadvantages. In the former approach, the probability density of the attributes is not always well fitted by a Gaussian distribution. In the latter approach, there can be a loss of information.

Instead of Gaussian distributions, we propose the use of fuzzy distribution graphs for the NBN classifier in the following manner. The continuous attributes are fuzzified and combined with probabilities of the naive Bayesian model in a simple fashion. The learned fuzzy distribution graphs are constructed by a neuro-fuzzy network called a neural network with weighted membership functions (NEWFM) [6, 7]. A NEWFM algorithm characterizes graphs of each feature using bounded sum of weighted fuzzy membership functions (BSWFM) [6, 7]. We compare the Gaussian distribution graphs with the fuzzy distribution graphs for the NBN classifier.

We applied the two types of distribution graphs to classify leukemia and colon DNA microarray data sets [8, 9]. With the successful completion of the Human Genome Project (HGP), we are entering the post genomic era. Facing massive amounts of data, traditional biological experiments and data analysis techniques encounter significant challenges. In this situation, cDNA microarrays and high-density oligonucleotide chips, which are novel biotechnologies, are global (genome-wide or system-wide) experimental approaches that are effectively used in the systematic analysis of large-scale genome data [14]. In this paper, we apply NBN classifiers for the analysis of DNA microarray data.

The experiment results demonstrate that NBN classifiers with fuzzy distribution graphs are more reliable than with Gaussian distribution graphs.

2. MATERIALS AND METHODS

2.1 Materials

In [8], the authors present methods for analyzing gene expression data obtained from DNA microarrays in order to classify types of leukemia. The data is split into two subsets: a training set and a test set. The training set consists of 38 samples (27 ALL and 11 AML) from bone marrow specimens. The test set has 34 samples (20 ALL and 14 AML), prepared under different experimental conditions and including 24 bone marrow and 10 blood sample specimens. All samples have 7,129 features, corresponding to some normalized gene expression value extracted from the microarray image. We used the same training and test set (as [8]) for our experiment. In [9], the authors describe and study a colon data set that is available on-line. Gene expression information was extracted from the DNA

microarray data resulting in a table of 62 tissues × 2000 gene expression values. The 62 tissues include 22 normal and 40 colon cancer tissues. Since there was no defined training or test set, we used all 62 samples for both training and test.

2.2 Preprocessing of The Data Set

In this section, we explain the data normalization and feature selection methods. We executed two steps for the data normalization. The first step was to normalize each input data to a scale of [0,1]. Next we provided a limited range of values using the Sigmoid method. In the feature selection, we used the Bhattacharyya distance method to rank the genes, and then selected the gene with the greatest distance as the best feature.

2.2.1 Data Normalization

To analysis microarray data, we normalized the microarray data in two steps. The first step, using formula (1), normalized each input value to the scale of [0,1]. Step 2 used the Sigmoid method to concentrate the data [10]. The Sigmoid formula is shown in formula (2). The matrix of genes is shown in Table 1. In this gene matrix, g_i is the *i*th gene. *p* and *q* represent class 1 and class 2 respectively. s_k and s_j are the *k*th sample in class 1 and the *j*th sample in class 2, respectively.

〈Table 1〉 Matrix of Genes

Genes	g_1	g_2	g_3	...	g_i
Samples in Class 1	$s_1^p(g_1)$	$s_1^p(g_2)$	$s_1^p(g_3)$	...	$s_1^p(g_i)$
	...	...	...	...	...
Samples in Class 2	$s_k^p(g_1)$	$s_k^p(g_2)$	$s_k^p(g_3)$	...	$s_k^p(g_i)$
	$s_1^q(g_1)$	$s_1^q(g_2)$	$s_1^q(g_3)$	...	$s_1^q(g_i)$
	...	...	...	...	...
	$s_j^q(g_1)$	$s_j^q(g_2)$	$s_j^q(g_3)$	...	$s_j^q(g_i)$

In formula (1), g_i is the original microarray value from the *i*th gene, $g_{\min}$ and $g_{\max}$ are the minimum value and maximum values from g_i, respectively. $g_i^{'}$ is the normalized value from Step 1.

$$f(g_i^{'}) = \frac{1}{1+e^{-g'_i}} \tag{2}$$

In formula (2), $f(g_i^{'})$ is the final normalized data that is used as the input data for our experiment.

2.2.2 Feature Selection

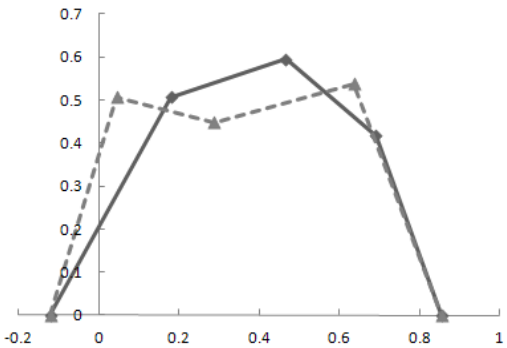
In this paper, we used the Bhattacharyya distance (BD) approach as our feature selection method for the microarray data analysis [11]. The definition of g_i's BD between two classes is shown as follows [11]:

$$B_i = \frac{1}{4}\frac{(\mu_i^p(s_k)-\mu_i^q(s_j))^2}{(\sigma_i^p(s_k))^2+(\sigma_i^q(s_j))^2} + \frac{1}{2}ln\left(\frac{(\sigma_i^p(s_k))^2+(\sigma_i^q(s_j))^2}{2\sigma_i^p(s_k)\sigma_i^q(s_j)}\right) \tag{3}$$

In formula (3), B_i denotes g_i's BD between two classes. $\mu_i^p(s_k)$, $\mu_i^q(s_j)$ are the g_i's mean value of genes for the samples of class1 and class2, respectively. $\sigma_i^p(s_k)$, $\sigma_i^p(s_j)$ denotes g_i's standard deviation of genes for the samples of class 1 and class 2, respectively. The gene with the greatest distance is the most differently expressed gene (DEG)[11].

2.3 Fuzzy Distribution Graphs for Naive Bayesian Classifier

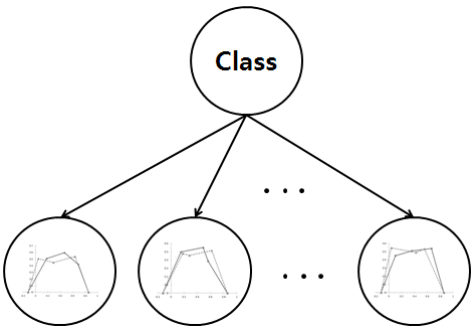
Instead of Gaussian distribution graphs, we proposed using fuzzy distribution graphs in the NBN classifier. We determined the membership function distribution from NEWFM [6, 7]. NEWFM can get BSWFM since train the experiment data [6, 7]. We applied the BSWFM into the NBN classifier. Fig. 1 shows an example of a fuzzy distribution graph for a leukemia gene. In Fig 1, the solid and dotted lines represent the fuzzy distrbiution of ALL and AML, respectively.



[Fig. 1] Fuzzy Distribution Graph for a Leukemia Gene

2.4 Constructing NBN Classifier Using Fuzzy Distribution Graphs

After building the fuzzy distribution graphs, we used these to construct the NBN classifier. The structure of the NBN classifier, using fuzzy distribution graphs, is shown in Fig. 2. The NBN classifier has one root node that represents the class and n leaf nodes that represent the features. Instead of Gaussian distributions, we used fuzzy distribution graphs on the leaf nodes.



[Fig. 2] Fuzzy Distribution Graph for a Leukemia Gene

Our process first normalized the microarray data. Then we used the BD to select the best features. With the selected features, the NEWFM algorithm characterized the fuzzy distribution graphs of each feature using BSWFM. Instead of Gaussian distribution graphs, fuzzy distribution graphs were applied to the NBN classifier. Finally we used these fuzzy distribution graphs with NBN classifier to classify the leukemia and colon DNA microarray data sets.

3. EXPERIMENTAL RESULTS

In this research, we selected the six best genes from the leukemia data set and the seven best genes from the colon data set [12, 13]. We compared the Gaussian distribution graphs (GDG) with the fuzzy distribution graphs (FDG) for the NBN classifier, and applied the two types of distribution graphs to classify the leukemia and colon DNA microarray data sets. The comparison results shown in Table 2, demonstrate that the NBN classifier with fuzzy distribution graphs was more reliable than with the Gaussian distribution graphs.

The performance results of this study showed that 100% accuracy could be achieved with a NBN classifier using fuzzy distribution graph for the leukemia data set. With the Gaussian naive Bayesian classifier, it could only attain 94%. For the colon data set classification, the accuracy of the fuzzy distribution graph NBN classifier was 8% higher than the Gaussian distribution graph.

〈Table 2〉 Experimental Results Comparison

Categories	Colon		Leukemia	
Method	GDG	FDG	GDG	FDG
Accuracy	81%	89%	94%	100%

4. CONCLUDING REMARKS

In this paper, we analyzed the data using fuzzy distribution graphs for the NBN classifier. We showed that the NBN classifier with fuzzy distribution graphs was more reliable than with Gaussian distribution graphs.

In future works, we will apply fuzzy distribution graphs to other Bayesian classifiers, such as the Tree-augment naive Bayesian (TAN) classifier. The NBN classifier assumes a strong independence, i.e., every attribute (every leaf node in the network) is independent of the other attributes. In the real world, nothing is absolutely independent. TAN approximates the interactions between attributes using a tree structure imposed on the NBN structure.

ACKNOWLEDGMENTS

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the IT-CRSP(IT Convergence Research Support Program) (NIPA-2013-H0401-13-1001) supervised by the NIPA(National IT Industry Promotion Agency).

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology. (2012R1A1A2044134).

REFERENCES

[1] N. Friedman, D. Geiger, and M. Goldszmidt, Bayesian network classifiers. Machine Learning, Vol. 29, No. 9, pp. 131-163, 1997.

[2] B. W. Morgan, An Introduction to Bayesian Statistical Decision Processes. New Jersey, 1968.

[3] D. Heckerman, A tutorial on learning with Bayesian Network. Microsoft Research, Redmond, 1996.

[4] M. Sahami, Learning Limited Dependence Bayesian Classifiers. Knowledge Discovery and Data Mining, pp. 335-338, 1996.

[5] P. Langley, An analysis of Bayesian classifiers. Tenth National Conference on Artificial Intelligence pp. 223 - 228, 1992.

[6] J. S. Lim, D. Wang, Y.-S. Kim, and S. Gupta, A neuro-Fuzzy Approach for Diagnosis of Antibody Deficiency Syndrome. Neurocomputing 69, Issues 7-9, pp. 969-974, March 2006.

[7] J. S. Lim, "Finding Features for Real-Time Premature Ventricular Contraction Detection Using a Fuzzy Neural Network System," IEEE Transactions on Neural Networks, pp. 522-527, 2009.

[8] T. R. Golub, D. K. Slonim, and P. Tamayo, Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, pp. 531-537, 1999.

[9] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. Proc. Nat. Acad. Sci. USA, vol. 96, pp. 6745-6750, 1999.

[10] H. Jun; M. Claudio, The influence of the sigmoid function parameters on the speed of backpropagation learning. From Natural to Artificial Neural Computation, pp. 195 - 201, 1995.

[11] G. Xuan, Bhattacharyya distance feature selection. Pattern Recognition, Proceedings of the 13th International Conference, 1996.

[12] X. W. Tian. S. H. Lee, and J. S. Lim, Gene Selection for Leukemia Classification Based on Bhattacharyya Distance. Proceedings of KIIS Spring Conference, vol. 23, No. 1, pp. 17-18, 2013.

[13] X. W. Tian and J. S. Lim, Bhattacharyya Distance for identifying differentially expressed genes in colon gene experiments, International Conference on Information Science and Applications, 2013.

[14] E. Themaat, On the Use of Learning Bayesian Networks to Analyze Gene Expression Data: Classification and Gene Network Reconstruction, Master Thesis, University of Amsterdam, Artificial Intelligence, June 2005.

전 설 위(Xue-Wei Tian)

- Aug. 2008 : Shandong University of Technology, China. Computer Science (B.S.)
- Aug. 2010 : Kyungwon University, Korea. Computer Science (M.S.)
- Mar. 2010 ~ now : Gachon University, Korea. Computer Science in Ph.D. course
- Interests : neuro-fuzzy systems, biomedical prediction systems, and signal process
- E-Mail : aitianxuemao@gmail.com

임 준 식(Joon S. Lim)

- 1986. 2 : Inha University, Korea. Computer Science (B.S.)
- 1989 : University of Alabama, Birmingham, USA. Computer Science (M.S.)
- 1994. 6 : Louisiana State University, USA. Computer Science (Ph.D)
- 1995. 3 ~ now : Gachon University, Korea. Dept. of Computer Science. professor.
- Interests : neuro-fuzzy systems, biomedical prediction systems, and human-centered systems
- E-Mail : jslim@gachon.ac.kr