A Nomogram Construction Method Using Genetic Algorithm and Naïve Bayesian Technique

KEON MYUNG LEE*, WON JAE KIM**, KEUN HO RYU*, SANG HO LEE*
*College of Electrical and Computer Engineering and PT-ERC
**College of Medicine
Chungbuk National University
Cheongju, Korea

{kmlee,wjkim,khryu,shlee}@cbnu.ac.kr

Abstract: - In medical practice, the diagnosis or prediction models requiring complicated computations are not widely recognized due to difficulty in interpreting the course of reasoning and the complexity of computations. Medical personnel have used the nomograms which are a graphical representation for numerical relationships that enables to easily compute a complicated function without help of computation machines. It has been widely paid attention in diagnosing diseases or predicting the progress of diseases. A nomogram is constructed from a set of clinical data which contain various attributes such as symptoms, lab experiment results, therapy history, progress of diseases or identification of diseases. It is of importance to select effective ones from available attributes, sometimes along with parameters accompanying the attributes. This paper introduces a nomogram construction method that uses a naïve Bayesian technique to construct a nomogram as well as a genetic algorithm to select effective attributes and parameters.

Key-Words: - nomogram, genetic algorithm, naïve Bayesian learning, medical data analysis, machine learning

1 Introduction[1]

In medical practice, it is sometimes hard to make decision for sure due to wide spectrum of individual variances in diseases and treatment effects. There have been tried to make use of computational models constructed from accumulated clinical cases in order to support such decision making. The task to build such computational models can be viewed as a machine learning problem.

In various medical domains, machine learning techniques have been applied to automatically build diagnosis models.[1] Although those models seem to slightly outperform the diagnosis accuracy, the technology has not be accepted widely in medical practice due to several reasons[1]: Inflexibility of the knowledge representation in which subjective information and fuzzy opinions are not easily incorporated in a formal and symbolic way[1]. Learning and classification techniques are sensitive to missing data which are often in the medical data[1]. The generated decision models typically include too few attributes which limit the explanation of decision[2]. Subjective resistance of physicians to new diagnostic technology even though they become aware that the technology is just not supportive and cannot replace the physicians[1]. Demanding pressure of evidence-based medicine has caused to develop computational models for medical decision making. When the decision making models are computationally complicated, it is not convenient for clinical personnel to use them. Hence, in clinical practice, monograms have been employed to alleviate such burden, which allow to get the model values by using graphical representation without conducting complicated number crunching. The construction of nomograms for diagnosis and progression has been tried over various diseases. It is challenging to discover an effective nomogram from clinical data set.

This paper is concerned with a nomogram construction method which uses genetic algorithm and naïve Bayesian technique. The proposed method employs a genetic algorithm to find proper predicting attributes and their parameters, and naïve Bayesian technique to determine a nomogram with respect to the selected attributes and parameters. The paper is organized as follows: Section 2 introduces nomograms

This work was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korean Government(MEST) through PT-ERC.

and notations used for the convenience of description. Section 3 presents a Bayesian-based nomogram construction method and Section 4 shows how to use a genetic algorithm to choose attributes and parameters for nomograms. Section 5 describes the procedures used to build a nomogram using genetic algorithm and naïve Bayesian learning, and Section 6 shows an application example of the proposed method. In final, Section 7 draws the conclusions.

2 Nomograms and Notations

A nomogram is a graphical representation to express numerical relationships that enables to easily compute a complicated function without help of computation machines. It is built to predict or diagnose a target such as disease development and recurrence, disease free survival. Figure 1 shows an example of nomogram. From a nomogram, attribute-wise scores are read according to their values and then added up to get the score sum. The probability to belong to the target class is determined by comparing the score sum with the reference probability values.

For the sake of description convenience, let us use the following notations:

$A = \{a_1, a_2, ..., a_m\}$: a set of attributes used to describe clinical data

$C = \{0,1\}$: the class label, where 0 indicates the other class and 1 the target class

$D = \{d_1, d_2, ..., d_n\}$: a set of patient data

$d_i = (v_{i1}, v_{i2}, ..., v_{im}, c_i)$: the *i*-th data, whether v_{ij} is the value for attribute a_j, and c_i is the class label

N_j: the *j*-th nomogram

$E_{N_j}(d_i)$: an evaluation function to measure whether data d_i belongs to the target class

$(d_{(1)}, d_{(2)}, ..., d_{(n)})$: the sorted sequence of the data set D in the increasing order of $E_{N_j}(d_i)$

$D_P = \{d_k | c_k = c, d_k \in D\}$: the subset of D, of which data belongs to the target class c

$D_N = \{d_k | c_k \neq c, d_k \in D\}$: the subset of D, of which data does not belong to the target class c

$n_P = |D_P|$: the number of data in D_P

$n_N = |D_N|$: the number of data in D_N

$p_j^P(v) = \frac{|\{d_i | a_i^j = v, d_i \in D_P\}|}{n_P}$: the relative frequency of data of which *j*-th attribute has value *v* in D_P

$p_j^N(v) = \frac{|\{d_i | a_i^j = v, d_i \in D_N\}|}{n_N}$: the relative frequency of data of which *j*-th attribute has value *v* in D_N

3 Naïve Bayesian-based Nomogram Construction

Nomograms can be regarded as a learned model because they are constructed to take the general picture in linear combination of individual scores from an accumulated set of clinical data. Hence, some machine learning techniques can be applied to build such nomograms. As a method to construct monograms, the naïve Bayesian models have been applied[3,4].

A naïve Bayesian classifier model is a simple probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumptions among events or attributes of data. Under the independence assumption on attributes, the posterior probability $P(c|X)$, the probability for an instance $X = (a_1, a_2, ..., a_m)$ to be a member of class *c*, is computed as follows:

$$P(c|X) = \frac{P(a_1, a_2, ..., a_m | c)P(c)}{P(X)} = \frac{P(c)\prod_i P(a_i | c)}{P(X)}$$

A nomogram can be considered as a model to evaluate how much an instance fits to a class. Let *c* be the class for which a nomogram is constructed, and $\bar{c}$ be the class other than *c*. $P(\bar{c}|X)$ denotes the probability that an instance *X* is not a member of class *c*. The odd ratio *Odds* for these two probabilities is defined and expressed as follows:

$$Odds = \frac{P(c|X)}{P(\bar{c}|X)} = \frac{P(c)\prod_i P(a_i | c)}{P(\bar{c})\prod_i P(a_i | \bar{c})}$$

logit is defined as logarithm of *Odds*. Hence, *logit* of $P(c|X)$ can be expressed as shown below:

$$\log it\, P(c|X) = \log it P(c) + \sum_i \log \frac{P(a_i | c)}{P(a_i | \bar{c})}$$

$$= \log itP(c) + \sum_i \log OR(a_i)$$

The above equation tells that *logit* of class probability $P(c \mid X)$ is determined by the sum of independent *logOR* of attribute values, $\log OR(a_i)$. This property enables a naïve Bayesian classifier model to be used in building a nomogram for which the final probability is obtained by adding up the individual evaluation of each attribute.

The following procedure shows how to build a nomogram based on the above observations.

1. Compute the relative frequencies, $p_i^P(v_j)$ and $p_i^N(v_j)$, of all possible values v_j of each attribute a_i with respect to the target class c and the other class $\bar{c}$ based on the given data set *D*.
2. Compute the *log OR,* $\log OR(v_j) = \log \frac{p_i^P(v_j)}{p_i^N(v_j)}$, for attribute values v_j over each attribute a_i.
3. Regard $\log OR(v_j)$ as the score of the corresponding attribute, and define the evaluation function $E(d_i)$ for data d_i as $\sum_j \log OR(v_{ij})$.

$$E_{N_j}(d_i) = \sum_{a^k \in SAT} \log OR(v_{ij}), \text{ where}$$

$$\log OR(v_{ij}) = \log_{10} \frac{p_j^P(v_{ij})}{p_j^N(v_{ij})}$$

4. Find the maximum *max* and minimum *min* among all possible attribute value combinations.
5. Over the range [*min*, *max*], determine the corresponding probability $p(c \mid d_i)$ for data d_i using the following equation:

$$P(c \mid X) = \left[1 + e^{-\log itP(c) - E(d_i)}\right]^{-1}$$

6. Draw the nomogram based on the score scales.

On constructing a nomogram from clinical data, the above procedure assumes that all attributes significantly influence the class of data. However, it is not always the case. Some attributes is irrelevant to the class information and further makes it difficult to identify class by introducing noises. Hence it is one of important issues to choose appropriate attribute set used in nomogram construction. In addition, many attributes are numeric and thus it is crucial to choose the right stratification to a numeric domain into categorical groups because the monograms normally handle categorical values.

4 A Genetic Algorithm-based Attribute and Parameter Selection Method

In order to choose the attributes to be used and the categorical quantization of numeric attributes, the proposed method makes use of a genetic algorithm-based procedure. Genetic Algorithms are a search algorithm based on the mechanics of natural selection and natural genetics. Motivated by the biological adaptation, they generate a new set of chromosomes from parent chromosomes via stochastic operations. A chromosome corresponds to a candidate solution for a given problem. Chromosomes with high fitness values survive and those with low fitness values die off generation to generation. While randomized, genetic algorithms efficiently exploit historical information to speculate on new search points with expected improved performance.

In order to use genetic algorithm approach to solve some problem, the following components of genetic algorithms should be developed: *Encoding Scheme* to code candidate solutions in chromosomes. *Genetic Operators* used to create new chromosomes from the existing chromosomes. *Fitness Evaluation Method* for all candidate solutions represented by chromosomes. *Population Initialization* to produce an initial pool of candidate solutions.

4.1 Candidate Representation

Binary coding is basically used for attribute selection and parameter selection. Parameter selection is needed when an attribute takes a value from a numeric domain. Such a domain is discretized into intervals and thus the boundary values for intervals have to be determined. The bit for an attribute selection is used to tell whether the attribute is included or not. For each numeric attribute, a specified number of bits are allocated and used to code a boundary value on the domain. If more intervals are needed, as many as boundary values have to be coded.

4.2 Operators

Bitwise crossover and mutation operators are used regardless whether a bit corresponds to an attribute or coding part of a parameter value.

4.3 Fitness function

Once a chromosome is given, a nomogram is constructed using the naïve Bayesian-based method as presented in Section 3. In order to determine the fitness of a nomogram, the next procedure is used:

1. For each data d_i of D, compute the evaluation value $E_{N_j}(d_i)$ with respect to nomogram N_j.
2. Sort the data set D in the increasing order of their score, along with the target label; $(d_{(1)}, d_{(2)}, ..., d_{(n)})$.
3. The fitness $fitness(N_i)$ of nomogram N_j is computed as follows:

$$fitness(N_i) = \Delta h \times \left(1 - \frac{DescH}{\Delta h}\right)$$

$p_k = \frac{\left|\left\{d_{(j)} \mid c_{(j)} = c, j = 1...k\right\}\right|}{k}$: the portion of data whose class label matches with class c in the first k data in $(d_{(1)}, d_{(2)}, ..., d_{(n)})$

$\Delta h = p_n - p_1$

$DescH = \sum_{i=1,(p_{i+1}-p_i)<0}^{n-1} (p_{i+1} - p_i)$: the accumulated sum of neighboring downward changes of p_i $(i = 1, ..., n)$

$SAT = \{a_{(1)}, a_{(2)}, ..., a_{(K)}\}, where\ a(k) \in A$: the set of attributes selected by the chromosome under consideration

The idea of the fitness function is to find the nomogram of which distribution of p_k over the sorted data $(d_{(1)}, d_{(2)}, ..., d_{(n)})$ is most likely to be monotonous and to maximize the height of the distribution.

5 Nomogram Construction

The proposed method finds a nomogram using the following procedure: As the input, the clinical data are given each of which is made of attribute values and class label, and the number of partitions for each numeric attribute is needed to be provided.

procedure *nomogram-construction*

1. Compute the relative frequencies, $p_i^P(v_j)$ and $p_i^N(v_j)$, of all possible values v_j of each attribute a_i with respect to the target class c and the other class $\bar{c}$ based on the give data set D.
2. Compute the *log OR*, $\log OR(v_j) = \log \frac{p_i^P(v_j)}{p_i^N(v_j)}$, for attribute values v_j over each attribute a_i.
3. Repeat the genetic algorithm introduced in Section 4 until the termination condition is satisfied, to search for nomograms as it selects effective attributes and, if needed, their parameters.
4. Choose the nomogram N with the highest fitness from the population of the genetic algorithm.

$$N = \arg\max_{N_i} fitness(N_i)$$

6 An Application Example

In order to see the applicability, the proposed method was applied to a real clinical nomogram construction problem with a bladder cancer patient data set for 166 patients. Each data contains demographic information, clinical information including whether the cancer recapped for the corresponding patient. Figure 1 shows a nomogram for bladder cancer recurrence obtained by the proposed method. The proposed method found a nomogram for bladder cancer recurrence which uses five attributes instead of available 10 attributes.

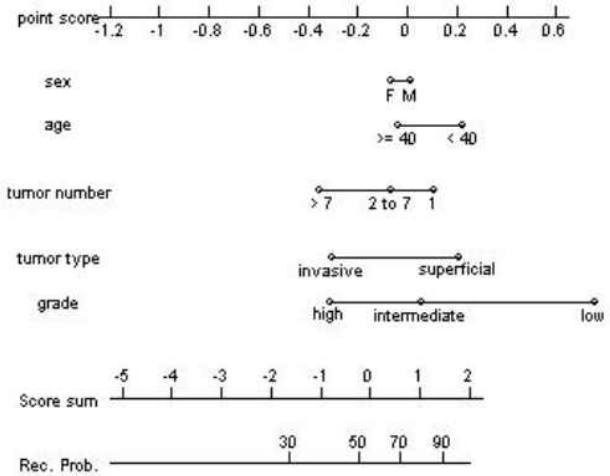

Figure 1. A nomogram for bladder cancer recurrence

Figure 2 shows a distribution to show the quality of nomogram by displaying p_k values over the sorted data sequence $(d_{(1)}, d_{(2)}, ..., d_{(n)})$. The quality of a nomogram is measured by the fitness function $fitness(N_i)$ with respect to the distribution like Figure 2.

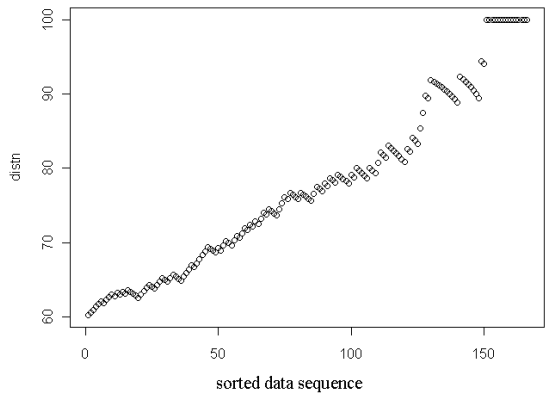

Figure 2. A distribution to show the quality of a nomogram performance

7 Conclusions

In medical practice, personalized medicine has been paid attention to provide better medical services. Nomograms are one of tools used for evidence-based medicine which exploits accumulated clinical experience. Nomograms would be better to evolve as clinical cases are added up. It takes time and effort to construct nomograms from available case data. The proposed method allows us to delegate to an information system the task to model a nomogram for a clinical data set with the help of genetic algorithm and naïve Bayesian learning technique. It has been applied to a real clinical problem and its result was evaluated as being worthy by a medical expert.

References:

[1] I. Kononenko, Inductive and Bayesian Learning in Medical Diagnosis, *Applied Artificial Intelligence*, vol.7, pp.317-337, 1993.

[2] V. Pirnat, I. Kononenko, T. Janc, I. Bratko, Medical Estimation of Automatically Induced Decision Rules, *Proc. of 2nd Europ. Conf. on Artifical Intelligence in Medicine*, pp.24-36, 1989.

[3] M. Mozina, J. Demsar, M. Kattan, B. Zupan, Nomograms for Visualization of Naïve Bayesian Classifier, *Proc. of PKDD 2004*, *LNAI 3202*, pp.337-348, 2004.

[4] M. Mozina, J. Demsar, M. Kattan, B. Zupan, Nomograms for Naïve Bayesian Classifiers and How can They Help in Medical Data Analysis, *Proc. of MEDINFO 2004*, pp.1762, 2004.

[5] A. Jakulin, M. Mozina, J. Demsar, M. Kattan, B. Zupan, Nomograms for Visualizing Support Vector Machines, *Proc. of SIGKDD'05*, 2005.