Reliability Index for 2-class Classification in Diagnosis of Disease: R_{I_{bin}}

Masahiro Nakatochi
h051315m@mbox.nagoya-u.ac.jp

Hiro Takahashi
h041305d@mbox.nagoya-u.ac.jp

Hiroyuki Honda
honda@nubio.nagoya-u.ac.jp

Department of Biotechnology, School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

Keywords: diagnosis, fuzzy classifier, reliability index

1 Introduction
According to development of bioinformatics, it has been possible to obtain large amount of gene expression data or single nucleotide polymorphism (SNP) data, and methods that extract significance genes from these exhaustive expression data have been proposed. Disease diagnosis models are constructed by their exhaustive data. It is important to discriminate reliability of individual estimate obtained from the diagnosis model. In our previous study, when fuzzy neural network (FNN) [1] is used to construct models, we proposed Reliability Index, R_I [2] which denotes reliable for continuous estimate. R_I can be calculated from feature of modeling data existing around diagnosis data. However, in diagnose of disease, estimate is binary values: benign or malign, or responder or non-responder, thus previous R_I for continuous estimate is inappropriate. In this study we developed Reliability Index for binary values.

2 Method
2.1 Reliability Index for Binary Values (R_{I_{bin}})
In this study, given a database from N people as the data set to construct the model (training data set), we have a set of gene expression profiles \( x^{tr}_{i} \) indexed by \( i \in \{1, 2, \ldots, N\} \). Each expression profile \( x^{tr}_{i} = [x^{tr}_{i1}, x^{tr}_{i2}, \ldots, x^{tr}_{in}] \) is a n-dimensional vector representing the measured expression levels of n genes. The class membership of each database sample is known and is denoted by \( f(x^{tr}_{i}) \). Diagnosis data estimated by constructed model is called test data. R_{I_{bin}} of the estimate of test data is defined in the form:

\[
R_{I_{bin}} = \sum_{i \in C} \left( Ra_{i} \times f \left( x^{tr}_{i} \right) \right) \quad (1)
\]

\[
Ra_{i} = - \log_{e} \left( \frac{e_{i}}{r_{i}} \right) \quad (2)
\]

In equation (1), when the coordinate of test data was regarded as the center in n-dimensional space, C is data set that exists nearest center in each quadrant of the space (total number of quadrants is \( 2^n \)). \( f(x^{tr}_{i}) \) is the class membership of each nearest training data that belong to C. In equation (2) \( e_{i} \) is the estimation error of each nearest training data that belong to C. \( r_{i} \) is the distance from test data to of each nearest training data that belong to C. \( Ra_{i} \) is the ratio \( \log_{e} \) and \( r_{i} \). If many training data that belong to the same class exist near test data and the estimation error of their training data is small, R_{I_{bin}} will be high, thus estimate of the test data is reliable.

2.2 Threshold of R_{I_{bin}} (R_{I_{thr}})
R_{I_{bin}} can show how the estimate of the test data is reliable. However it isn’t clearly how R_{I_{bin}} of reliable estimate is high, thus we propose threshold for R_{I_{bin}} (R_{I_{thr}}) in the form:

\[
R_{I_{thr}} \equiv \sum_{i=1}^{n} \left( Ra_{av} \right) \times f \left( x^{av}_{i} \right) = 2^n \times Ra_{av} \quad (3)
\]

\[
Ra_{av} = - \frac{\log_{e} e_{av}}{r_{av}} \quad (4)
\]
In equation (4), $e_{av}$ is the estimation error average of all training data. $r_i$ is the distance average from test data to each nearest training data that belong to $C$. If one estimate has higher $R_{bin}$ than $R_{thr}$, its estimate is reliable. Otherwise, its estimate is unreliable.

2.3 Data Set
In this study, we used multiple simulation data sets made randomly and microarray data set. As an example of simulation data set, following function was given: If a data have all expression value of appointed genes ($n=3$) is higher than specific value, its data belongs to $f(x^{(tr)}) = +1$, otherwise $f(x^{(tr)}) = -1$. As an actual diagnosis, we also used data set of breast cancer patients from a study of van’t Veer L.J et al[3]. The data set comprised 24,481 genes and 97 patients. These 97 patients are assigned 2-class; one class composed of 46 patients that developed distant metastases within 5 years, and another class composed of 51 patients that continued to be disease-free after a period of at least 5 years. In this experiment, the data set was randomly partitioned into two groups: 78 patients as training data and 19 patients as test data set. We excluded those genes that compromise missing value in 97 patients. Thus, 13,547 genes remained. We selected 3 genes from 13,547 genes by using t-test. These 3 genes was used to construct the model ($n=3$).

![Graph](image1.png)

Figure 1 example of distributions of $R_{bin}$ in multiple simulation data set.

![Graph](image2.png)

Figure 2 Accuracy of estimates selected by $R_{thr}$

3 Results and Discussion
At first we confirmed that there are many incorrect data in low $R_{bin}$ and many correct data in high $R_{bin}$ by using multiple simulation data sets made randomly. In Figure 1, we show example of distributions of $R_{bin}$ in multiple simulation data sets. Two distributions are clearly separated for $R_{bin}$. Furthermore, we confirmed availability of $R_{thr}$ and $R_{bin}$ for microarray data set of breast cancer patients. In Figure 2, the left bar is “unselected” accuracy of all estimates, and the right bar is “selected” accuracy of only estimates that has higher $R_{bin}$ than $R_{thr}$. “Unselected” accuracy is 52.6%. By using $R_{bin}$ and $R_{thr}$, “selected” estimate correspond to 40% of all patients, and “selected” accuracy rose until 76.7%. These results show $R_{thr}$ and $R_{bin}$ are available for estimation in binary classes.

References
