Implementation of mutual information and bayes theorem for classification microarray data

Mahendra Dwifebri Purbolaksono, Kurnia C Widiastuti, Mohamad Syahrul Mubarok, Adiwijaya, Firda Aminy Ma’ruf

School of Computing, Telkom University

Abstract. Microarray Technology is one of technology which able to read the structure of gen. The analysis is important for this technology. It is for deciding which attribute is more important than the others. Microarray technology is able to get cancer information to diagnose a person’s gen. Preparation of microarray data is a huge problem and takes a long time. That is because microarray data contains high number of insignificant and irrelevant attributes. So, it needs a method to reduce the dimension of microarray data without eliminating important information in every attribute. This research uses Mutual Information to reduce dimension. System is built with Machine Learning approach specifically Bayes Theorem. This theorem uses a statistical and probability approach. By combining both methods, it will be powerful for Microarray Data Classification. The experiment results show that system is good to classify Microarray data with highest F1-score using Bayesian Network by 91.06%, and Naïve Bayes by 88.85%.

1. Introduction

Cancer is the leading cause of death for everyone around the world. According to WHO’s data, there are about 14 million new cases just in 2012 which can increase next year. Early observation of disease is needed to prevent cancer growth. Furthermore, cancer detection is also important. There are several ways to detect cancer, such as physical detection, blood test and X-ray. But these could take a long time and containing human errors during detection. In 2013, British Medical Journal (BMJ) released that medical error caused 251,454 death every year in United States of America. Therefore, it is required to have technologies that can solve the problem. One of them is Gene Expression Technology that involves Microarray technology for cancer detection [1].

Microarray technology is a developing technology used to study the expression of many genes at once [2]. It is the groundwork for many cancer detection researches. However, the main problem of Microarray data is processing the data itself since it contains many number of insignificant and irrelevant attributes [2] which affect the classification process and result. Therefore, it needs a method to reduce the dimension of Microarray data without eliminating important information in every attribute. After reducing the dimension, the data will be classified based on the existing classes.

The problem of data classification will be solved using Bayesian methods. For classifying Microarray data, the methods have attracted a lot of attentions from researchers to study this topic, such as Liying Yang [3], Devi Arockia [4] and Luis M. de Campos [5]. In reference [5] it explained
the detail about Microarray, such as data distribution and approach to solve Microarray classification problems. Many different approaches have been used, one of them is feature selection approach using Mutual Information (MI) and Support Vector Machine (SVM) for the classifier [4]. In the research, it showed that MI has good performance for dimensional reduction. However, for classification process, this research has a deficiency for classifying multi-class data. While in research [5], it used a variety of Bayes Theorem such as Naïve Bayes and Bayesian Network that have good performance in this topic. But in research [5], it has deficiency for data dimensional reduction process. That happened because Certainty Factors method can only process maximum 2 data. If it is more than two data, the system needs more time to execute it.

Therefore, this research uses Mutual Information for feature selection and it is expected to produce a better result after dimension gets reduction. While for classification, it uses Naïve Bayes and Bayesian Network since they have ability to get pattern and dependency among variables using probability approach.

2. Study of Literature
Two main focuses in this research are the usage of Mutual Information and Bayes Theorem. The Bayesian Theorem that we used are Naïve Bayes and Bayesian Network. For the system, there are three main stages namely Discretization, Feature Selection and Learning of Bayesian.

2.1. Discretization
We used K-Means algorithm to transform continue data into discrete data. K-Means works by dividing data into several groups, also called clustering. Firstly, K-Means randomly distribute center points (centroids) for clusters representation. Because of randomly distribution, it is possible that some centroids are not properly located. Therefore, they need to be updated as follows,

\[
\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q
\]

where \( \mu_k \) is centroid of \( k \)th cluster, \( N_k \) is number of data belongs to \( k \)th cluster, and \( x_q \) is \( q \)th data on \( k \)th-cluster. Updating centroids is done until centroids do not change.

2.2. Feature Selection
The problem in Microarray data is about large dimension of the gene. Therefore, we need to reduce the number of dimension. One of the solution is using Mutual Information. In this work, Mutual Information is used to select informative gene from the original gene expression profile [6]. To compute Mutual Information, the probability distribution of genes is needed which in practice are not known, and the best we can do is using histogram of data. The steps involved in computing the Mutual Information from the histogram of the training data are given as follows,

a. The data set is arranged in an ascending order based on the output.

b. The output class label (Y) is divided into two groups and the initial entropy \( H(Y) \) is calculated using

\[
H(Y) = - \sum_{i=1}^{N_y} P(Y=i). \log(P(Y=i))
\]

(2)

c. The input genes (X) are divided into ten levels and their conditional entropies \( H(Y|X) \) are evaluated using

\[
H(Y|X) = - \sum_{i=1}^{N_x} P(X=i). \sum_{j=1}^{N_y} P(Y=j|X=i). \log(P(Y=j|X=i))
\]

(3)

d. Next, the mutual information of each gene with respect to the output is computed using

\[
I(Y; X) = H(Y) - H(Y|X)
\]

(4)

Where:

X = [Input number-i \{ X_1: 2.5, X_2: 1.5, ..., X_n: 2.25 \}]

Y = [Class number-i \{ Y_1: ALL, Y_2: AML \}]

e. The result of Mutual Information will be sorted ascending. First attribute is the highest value and will be selected as the most informative gene.
2.3. Bayes Theorem

Bayes Theorem is a theorem which is used in statistic for calculating probability of hypothesis. Bayes Optimal Classifier [7] calculates probability from each group of an existing class and for deciding the most optimal class. Naïve Bayes is the first considered as Bayesian Classifier.

Naïve Bayes or can be called as Bayes Theorem with the assumption that the attribute independency is a classification that depends on two assumptions. Those assumptions are as follow: each attribute will be independent given class attribute and all attributes have influence to a class. Assumption of attribute independency will eliminate the needs of amount training data from multiplication of Cartesian by all of attribute which used to classify a data [8]. In a Directed Acyclic Graph (DAG) which represents the Naive Bayes (NB) structure, all of the bows will be aimed from attribute class to all attributes like Figure 1.

![Figure 1. Example of Naïve Bayes Structure](image)

Where:

\[ C = \text{class} \{ \text{Leukemia, ALL, AML} \} \]

\[ x_1 = \text{attribute number-1} \{ x_1 : 250 \} \]

The development of Naïve Bayes method is accomplished by using many methods [13]. One of it begins from the basic topology, for example by connecting each attributes to one another. This method is one of the Bayes theorem methods which is also knows as Bayesian Network. For an example, Bayesian Network with a connection between each attribute like Figure 2.

![Figure 2. Example of Bayesian Network Structure](image)

Where:

\[ C = \text{class} \]

\[ x_n = \text{attribute} \]

From all of the methods, Bayes Theorem is using this equation [9]:

\[
P(A | B) = \frac{P(B | A)P(A)}{P(B)}
\]  

(5)

Where:

- A is data hypothesis A (specific class)
- B is data with unknown class
- P(A | B) is probability hypothesis A given B condition (posterior | probability)
- P(B | A) is probability B given hypothesis A condition
- P(A) is probability of hypothesis A (prior probability)
- P(B) is probability of B
The first step of this schema is to establish the Naive Bayes or Bayesian Network structure using MAP estimation (Maximum A Posterior).

To calculate the MAP, given \( r_i \) which signifies cardinality from \( X_i \), and \( q_i \) that represent cardinality from parent set \( X_i \). After that, the conditionally probability \( P(X_i|pa(X_i)) \) can be represented into

\[
\theta_{ijk} = \frac{N_{ijk} + \alpha_{ijk}}{N_{ij} + \alpha_{ij}}
\]

(6)

Where \( N_{ijk} \) is the amount of data in \( D \) for \( X_i \) obtained from \( k \) value and parent \( pa(X_i) \) which is obtained from \( j \) value. \( N_{ij} \) is the amount of data from \( D \) for each \( pa(X_i) \) obtained from \( j \) value. With \( \alpha_{ijk} \) and \( \alpha_{ij} \) value as follows.

\[
\alpha_{ijk} = \frac{\alpha}{r_{ijk} q_{ijk}}
\]

(7)

\[
\alpha_{ij} = \frac{\alpha}{r_{ij} q_{ij}}
\]

(8)

Where \( \alpha \) value is a hyper parameter as a number which works for smoothing or avoiding being divided by zero. In this system \( \alpha \) value will be 0.001.

3. System Design

The main system flow is divided into two, which are training and testing. On the training flow, the input data is all of training data. The purpose of this process is to obtain the parameter from NB or BN based on the distribution of training data. The flow will be explained in Figure 3.

On the testing flow, it is more or less the same with training flow. The difference is the learning process of Bayes. That process is not required, but what is needed is the retrieval of previously trained Bayes model data to be interfered using joint probability for existing data testing. The form of testing input data is the same with training input data.

![Figure 3. Flow of System Design](image)

The system performance can be measured when the testing schema is done. Which means the model performance can be used for the new data, in this case the testing data. The model performance will be measured using F1-score.

\[
Precision = \frac{True \ Positive}{Total \ Predicted}
\]

(9)

\[
Recall = \frac{True \ Posative}{Total \ Target}
\]

(10)

\[
F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}
\]

(11)
4. Testing and Analysis
The dataset used is Kent Ridge Bio-medical [11], which is consisting of training data and testing data.

**Table 1. Distribution of dataset [11]**

| No | Data sets       | Number of Instances | Number of Genes | Number of Classes |
|----|-----------------|---------------------|-----------------|-------------------|
| 1  | Colon Cancer    | 62                  | 2000            | 2                 |
| 2  | Ovarian Cancer  | 253                 | 15154           | 2                 |
| 3  | Leukemia ALL-AML| 72                  | 7129            | 2                 |
| 4  | Breast Cancer   | 97                  | 24481           | 2                 |
| 5  | Lung Cancer     | 181                 | 12533           | 2                 |

The objective is to classify the microarray data accurately using a system that is built by combining different algorithms. To accomplish that objective, there are two scenarios used, which are the observation on the effect of dimensional reduction and the second scenario is the observation on the effect of K value in K-Means.

4.1. Scenario 1: Dimensional Reduction
In this scenario, the feature selection will be done using Mutual Information. Mutual Information method is one of the methods that is useful to reduce the dimension of microarray data. Therefore, all of the attributes can be observed by looking at the importance of each attribute. This scenario is important because it will solve the problem of microarray data which is its high dimension.

It has been observed that Colon Cancer with K=10, Ovarian Cancer with K = 8, Leukemia ALL-AML with K=10, Breast Cancer with K=10, and Lung Cancer with K=12. There are six K-Means combinations for each data with conditions mentioned in Table 2.

**Table 2. Distribution of MI value in every data sets**

| Combination | Colon Cancer          | Ovarian Cancer        | Leukemia ALL-AML | Breast Cancer | Lung Cancer |
|-------------|-----------------------|-----------------------|-----------------|--------------|------------|
|             | Range of MI Score     |                       |                 |              |            |
| 1           | 0.6544-0.6353         | 0.6541-0.6394         | 0.6016-0.6017   | 0.6849-0.6742| 0.6931-0.6931|
| 2           | 0.6544-0.6274         | 0.6541-0.6364         | 0.6016-0.5973   | 0.6849-0.6703| 0.6931-0.6873|
| 3           | 0.6544-0.6203         | 0.6541-0.6311         | 0.6016-0.4587   | 0.6849-0.6679| 0.6931-0.6817|
| 4           | 0.6544-0.6130         | 0.6541-0.6278         | 0.6016-0.3870   | 0.6849-0.6655| 0.6931-0.6814|
| 5           | 0.6544-0.6055         | 0.6541-0.6256         | 0.6016-0.3614   | 0.6849-0.6633| 0.6931-0.6758|
| 6           | 0.6544-0.3961         | 0.6541-0.6215         | 0.6016-0.3407   | 0.6849-0.6617| 0.6931-0.6752|

Those values are referring to the range of MI score which is every attribute score. Therefore, based on the value distribution, those results will be used in first scenario with five datasets and six combinations (1-6) such as Table 3.

**Table 3. Comparison Performance of Various Attributes Number**

| Combination | Colon Cancer | Ovarian Cancer | Leukemia ALL-AML | Breast Cancer | Lung Cancer |
|-------------|--------------|----------------|------------------|--------------|------------|
|             | NB           | BN             | NB               | BN           |            |
| 1           | 0.6667       | 0.7333         | 0.9677           | 0.588s^2     | 0.7647     |
|             |              |                |                  | 0.5789       | 0.5789     | 0.9731      | 0.9731     |
| 2           | 0.7333       | 0.7333         | 0.9677           | 0.9753       | 0.7647     | 0.7368      | 0.7368     | 0.9799      | 0.9801     |
| 3           | 0.7333       | 0.7333         | 0.9677           | 0.7647       | 0.7368     | 0.7368      | 0.9799      | 0.9801     |
| 4           | 0.7333       | 0.7333         | 0.9677           | 0.8823       | 0.7895     | 0.7895      | 0.9799      | 0.9801     |
| 5           | 0.7333       | 0.7333         | 0.9677           | 0.8823       | 0.7895     | 0.7895      | 0.9799      | 0.9801     |
|             | 0.8823       | 0.8823         | 0.8823           | 0.8420       | 0.9866     | 0.9866      | 0.9876     |            |            |
The overall analysis obtained from analyzing each of the datasets:

a. Increasing the range value of MI does not necessarily improve the performance result, this is shown by the performance results from using Ovarian Cancer and Leukemia dataset, where the performance results were not affected by the increment of the value range.

b. There is a max point in some of the dataset, like Breast Cancer, Colon Cancer and Lung Cancer. At first the results increased until it reached max point but then it decreased.

The analysis mentioned above are caused by different amount of attributes which become a unique characteristic of each datasets and also the diverse data distribution.

4.2. Scenario 2: Distribution Value of K in K-Means

Scenario 2 is done by clustering the data using K-Means. K-means method is one of methods that can be used for discretization. Thus all values on attributes that originally have a large distribution of data can be minimized. This scenario will be able to solve the problem of calculating Mutual Information, which is the different distribution between variables to class. The performance results comparison of the various K values is written at Table 4.

| Evaluation     | Colon Cancer | Ovarian Cancer | Leukemia ALL-AML | Breast Cancer | Lung Cancer | AVG |
|----------------|--------------|----------------|------------------|--------------|-------------|-----|
| K=2            | NB 0.6667    |                |                  |              |             |     |
| K=8            | FN 0.684     | 0.8            | 0.9839           |              |             |     |
| K=10           | FN 0.9508    | 0.7941         | 0.8823           | 0.8420       | 0.9866      | 0.9876 |
| K=12           | FN 0.950     | 0.765          | 0.984            | 0.984        | 0.631       |     |
| K=16           | FN 0.950     | 0.765          | 0.984            | 0.984        | 0.631       |     |

From all of the analysis of each dataset, it can be concluded that increasing the K value will not necessarily improve the performance result. However, each of dataset will have its own pattern of value. In Ovarian, Lung Cancer, Breast Cancer and Colon Cancer data sets have maximal score when K value is 10, but it did not apply in Leukemia by Bayesian Networks. These results occurred because every dataset has different characteristics or data distribution to each other.

4.3. Comparison of Maximal Average Result

After all of the experiments are finished, we got the maximum result of every data sets in Table 5.

| Evaluation     | Colon Cancer | Ovarian Cancer | Leukemia ALL-AML | Breast Cancer | Lung Cancer | AVG |
|----------------|--------------|----------------|------------------|--------------|-------------|-----|
| Max Accuracy   | NB 0.8       | 0.939          | 0.8823           | 0.7895       | 0.9866      | 0.8885|
|                | BN 0.867     | 0.984          | 0.882            | 0.84         | 0.980       | 0.9106|
Based on the results shown in Table 5, it proves that Mutual Information for feature selection method and Bayesian for classifier method can classify microarray data with the average score of Naïve Bayes is 0.8885 and the Bayesian Network is 0.9106. The results were obtained from the analysis of dimensional reduction and the distribution value of K in K-Means in every data set. In this research, Bayesian Network has the biggest average because the structure of Bayesian Network allows connection between each and every attributes.

5. Discussion and Conclusion

In microarray case, the first thing that should be noticed is the type of the data, whether it is continuous data or discrete data, both of them have wide data distributions. Therefore, clustering is required to solve that problem. This research used K-Means for discretization or clustering method. Amount of K for K-Means has different trend in every data, depends on the amount of random attribute or data sets type.

For feature selection, this research used Mutual Information method. From the feature selection process, it can be concluded that increasing the value range of MI will increase the number of random attribute. However, it does not guarantee that the performance will increase as well. Each data set has its own boundary to get its maximum F1-score. It would be better if the next observation focuses more on using the other discretization methods and feature selection methods. It will gain a new trend of data which can improve the performance.

The machine learning approach using Bayes Theorem has proven to successfully classify microarray data with good performance results. Based on one of the Bayes Theorem methods used which is the Bayesian Network, it is proven to acquire better performance in classifying the microarray data than Naïve Bayes method. Bayesian Network obtained 0.9106 average score. That score is good for classification of microarray data and used to detect cancer. Bayesian Network can be improved again with optimized DAG by Learning Structure. The future research itself is not limited in exploring the use of Learning Structure method only, but it can also be further expanded by looking into the other methods of feature selection, because it will help to gain new attributes.
References

[1] Nurfalih, Adiyasa; Adiwijaya; Suryani, Arie Ardiyanti, "Cancer Detection Based On Microarray Data Classification Using PCA and Modified Backpropagation," *Far East Journal of Electronics and Communications*, vol. 16, no. 2, pp. 269-281, 2016.

[2] Singh, Rabindra Kumar; Sivabalakrishman, Dr. M., "Feature Selection of Gene Expression Data For Cancer Classification: A Review," in *in 2nd International Symposium on Big Data dan Cloud Computing (ISBCC’15)*, Vellore, India, 2015.

[3] Liying Yang; Zhimin Liu; Xiguo Yuan; Jianhua Wei; Junying Zhang, "Random Subspace Aggregation for Cancer Prediction with Gene Expression Profile," *BioMed Research International*, vol. 2016, p. 10, 2016.

[4] C, Devi Arockia Vanitha; D, Devaraj; M, Venkatesulu, "Gene Expression Data Classification using Support Vector Machine and Mutual Information-based Gene Selection," in *Graph Algorithms, High Performance Implementations and Application (ICGHA2014)*, Tamil Nadu, India, 2015.

[5] Luis M. de Campos; Andres Cano; Javier G. Castellano; Seraffin Moral, "Bayesian Networks Classifiers for Gene-Expression Data," in *Department of Computer Science and Artificial Intelligence. Granada University.*, Spain, 2011.

[6] C, Devi Arockia Vanitha; D, Devaraj; M, Venkatesulu, "Gene Expression Data Classification using Support Vector Machine and Mutual Information-based Gene Selection," in *Graph Algorithms, High Performance Implementations and Application (ICGHA2014)*, Tamil Nadu, India, 2015.

[7] T. M. Mitchell, Machine Learning, New York: McGraw-Hill, 1997.

[8] Berson, A; J, Smith S., Data Warehousing, Data Mining, & OLAP, New York: McGraw-Hill, 2001.

[9] Almira Syawli; M. Ali Fahmi; Silvia Ari Santhy; Zulkarnaen, "Diagnosa Penyakit Diabetes Mellitus Berbasis Desktop Application," *Jurnal Diabetes Mellitus*, vol. 3, no. 2, pp. 50-62, 2010.

[10] Yun Zhou; Norman Fenton; Martin Neil, "Bayesian Network Approach to Multinomial Parameter Learning Using Data and Expert Judgments," *International Journal od Approximate Reasoning*, vol. 55, pp. 1252-1268, 2014.

[11] Bio-medical, Kent Ridge, "Kent Ridge Bio-medical Dataset," 2003. [Online]. Available: http://datam.i2r.a-star.edu.sg/dataset/krbd. [Accessed 12 2016].

[12] W. H. Organization, "Cancer fact sheet," Februari 2015. [Online]. Available: http://www.who.int. [Accessed 1 November 2016].

[13] Mubarok, MS; Adiwijaya; Aldhi, MD; “Aspect-based sentiment analysis to review products using Naive Bayes”. *In AIP Conference Proceedings* 1867, 020060, 2017.