The Research of Missing Value Estimation of Gene Sequence Based on Improved KNN

Cai Qing, Wu Qingfeng, Dong Huailin and Liu Han
Software School
Xiamen University
Xiamen, Fujian Province 361005, China
helen99ok@yahoo.com.cn

***Abstract*—Gene based data mining has been received wider and wider attention as gene carries genetic information of living creature. While mining gene information, one of the tasks is to estimate the missing values reasonably and effectively, so as to reflect the original information of gene sequence. By analyzing the theory of KNN (K nearest neighbor algorithm), an improved KNN for gene sequence was proposed, which resolves the problem of missing values while mining gene data. Results show the feasibility of the algorithm with experiments using data from genbank.**

***Index Terms*—Gene sequence, Missing values, KNN**

I. INTRODUCTION

DNA in living cells is inherently unstable. It is subject to mechanical stress and many types of chemical modification that may lead to breaks in one or both strands of the double helix [6]. Double helix strands also tend to break while recombination taking place between paired homologous chromosomes or rearrangement of immunoglobulin gene segments going on [6]. If the double-strand breaks can not be repaired in cells, they may cause not only death of a cell but also gene deletion [6]. It is found out that some types of gene deletion relate closely to some diseases, such as Loss of heterozygosity (LOH) affecting at least one locus of FHIT gene was observed in some tumors [3], the positive rates of FHIT gene deletion and HPV-DNA fragment in breast tumor cells were also respectively high [4]. Medical scientists have tried some gene therapy to treat the disease concerning deletion, for example, Salima and other researchers treated five patients suffering X-linked severe combined immunodeficiency due to a deficiency of the γc chain with autologous CD34+ cells from bone marrow that had been transduced ex vivo with the γc gene [7]. Replacing a defective gene with a normal gene and thus restoring the lost gene function in the patient's body is the essence of gene therapy, therefore, it is crucial to find the original gene corresponding to gene containing missing value in treatment of gene deletion-related disease.

By now, there are some prevalent methods in dealing missing values [5].

1) Delete the samples containing missing attributes from our dataset and get a new dataset. This method is simple, but cannot be applied here since we hope to estimate the missing values instead of neglect them.

2) Fill out the missing value manually. This method is time-consuming and relies heavily on the experience of staffs.

3) Replace missing values with a global value or an average value. However, the average value is hard to compute for a nonnumeric type data like gene data, and it is also not reliable to fill the missing value with a particular base (such as guanine). If we do apply this way to fill up missing values, the results will tend to loss reliability.

4) Estimate the missing value with a prediction model. A prediction model can be set up with other existing phrases of gene to predict the value of missing data. This method exploit the maximum knowledge of relevant data, is one of the prevalent methods in handling missing values, but still be subjected to the difficulty of predicting the distributed model of gene data.

At present, there isn't much literature about estimation of missing value in gene sequence, and most of the nonnumeric data based missing value estimation work is also done by replacing the missing value with the value shows up most frequently. With its simplicity, KNN has been widely used in many fields such as text classification. Based on the implementation of KNN, an improved KNN to estimate the missing values in genes is proposed, afterwards, the experiment will be introduced and the results will be analyzed.

The following of this paper will firstly explain the basic idea of KNN, later, a new distance computing method based on gene data will be proposed which followed by its detail utilization in improved KNN, at last, experiment and results will be illustrated and discussed.

II. KNN AND RELATING DISTANCE COMPUTING METHODS

With its efficiency and readiness of implementation, KNN has been widely used in many fields, such as classification and pattern recognition, it is recognized as a conventional and statistic based pattern recognition method. When applied to missing value estimation, KNN shows a robust and sensitive performance, its basic idea is to find samples that most similar to the sample containing missing values, and fills the missing

attributes in the sample with the known attribute values of the neighbor samples.

There are 4 steps in KNN method:

Step1. Load the dataset D containing missing values, suppose D is a $(n+1)\times m$ matrix with (n+ 1) sample and m attributes for every sample.

Step2. Initialize samples containing missing values, one of the common ideas is to fill the missing attribute with the mean of values of other attributes in the sample, for example, if the m_{th} attribute of sample a is missing, then the initial estimated value of a_m can be obtained from equation (1):

$$a_m = \frac{1}{m-1}\sum_{j=1}^{m-1} a_j \tag{1}$$

step3. After initializing all the missing values, the distance between sample a and other samples (denote them as $\{b_1, b_2 ... b_n\}$, and the j_{th} attribute value for gene b_i is expressed as $b_i^{(j)}$) can be computed. The distance between two samples is computed as Euclidean distance which shows in equation (2).

$$d(a,b_i) = \sqrt{\sum_{j=1}^{m}(a_j - b_i^{(j)})^2} \tag{2}$$

Step4. Sort the distances yielded in step (3), suppose $d(a, b_1)<d(a,b_2)<\ldots<d(a,b_n)$, then the p_{th} attribute in sample a which is missed (Denote it as a_p) can be estimated with equation (3).

$$a_p = \frac{1}{k}\sum_{i=1}^{k} b_i^{(p)}, k \le n \tag{3}$$

Dudani proposed a improved scheme which suggested that there should be a distance-weighted value for each neighbor [1], following this, the weight for sample b_i (denote it as w_i)can be figured out with equation (4).

$$w_i = \frac{1}{d(a,b_i)^2} \tag{4}$$

After the calculation of distance-weighted values for samples, the estimated value for a_p can be calculated with equation (5).

$$a_p = \frac{1}{kW}\sum_{i=1}^{k} w_i b_i^{(p)}, \qquad W = \sum_{j=1}^{k} w_j \tag{5}$$

III. THE IMPROVED KNN BASED ON GENE DATA

Although works fine for numeric data, Equation (2) has difficulty in getting distance between genes due to the difference between gene data and numerical data. Researchers proposed that the $(a_j - b_i^{(j)})$ in equation (2) can be handled as in equation (6) [2], such method has been adopted in text comparison and questionnaire survey.

$$(a_j, b_i^{(j)}) = \begin{cases} 1, & a_j \ne b_i^{(j)} \\ 0, & a_j = b_i^{(j)} \end{cases} \tag{6}$$

It is feasible to find the genes with high similarity with equation (2) and equation (6), however, the structural characteristics of DNA was neglected. As it known to all, DNA is nucleotide chains with double helix structure, and corresponding bases in the same place in the two chains complement one another. In gene sequencing, the whole sequence can be figured out by working out the sequences in one chain, therefore, if gene a and b_i are similar with each other, but we only get sequences from different chains, then the distance between a and b_i computed with equation (2) will loss veracity.

A modification is made on the distance computing method in KNN.

step1. Load the dataset D containing missing values.

Step2. Compute distance between gene a and other genes (denote them as $\{b_1, b_2, \ldots, b_n\}$, the length of a and b_i is expressed as m_a and m_i) in the dataset with equation (7), the j_{th} value for gene bi is expressed as $b_i^{(j)}$

$$d(a,b_i) = \max\left(\left|\sum_{j=1}^{\min(m_a,m_i)} d_+(a_j, b_i^{(j)})\right|, \left|\sum_{j=1}^{\min(m_a,m_i)} d_-(a_j, b_i^{(j)})\right|\right) * \gamma \tag{7}$$

Equation (8) and equation (9) explain the meaning of $d_+(a_j, b_i^{(j)})$ and $d_-(a_j, b_i^{(j)})$ in equation (7).

$$d_+(a_j, b_i^{(j)}) = \begin{cases} 1, & a_j = b_i^{(j)} \\ 0, & a_j \ne b_i^{(j)} \end{cases} \tag{8}$$

$$d_-(a_j, b_i^{(j)}) = \begin{cases} -1, & a_j = pair(b_i^{(j)}) \\ 0, & a_j \ne pair(b_i^{(j)}) \end{cases} \tag{9}$$

$pair(b_i^{(j)})$ means the base complementary to $b_i^{(j)}$, suppose $b_i^{(j)}$ is T (thymine), then $pair(b_i^{(j)})$ is A(adenine). Coefficient γ is -1 when $\left|\sum_{j=1}^{\min(m_a,m_i)} d_+(a_j, b_i^{(j)})\right|$ greater than $\left|\sum_{j=1}^{\min(m_a,m_i)} d_-(a_j, b_i^{(j)})\right|$, otherwise 1. Concerning missing attribute a_p, we assign 0 to both $d_+(a_j, b_i^{(j)})$ and $d_-(a_j, b_i^{(j)})$.

step3. Sort the distances yielded in step (2), suppose $|d(a,b_1)|>|d(a,b_2)|>\ldots>|d(a,b_n)|$, then the weight for each neighbor (the weight of neighbor b_i is denoted as w_i)can be computed using equation (10), but w_i should set to be 1 when $d(a,b_1)$ equals $d(a,b_k)$.

$$w_i = \frac{|d(a,b_1)| - |d(a,b_i)|}{|d(a,b_1)| - |d(a,b_k)|} \tag{10}$$

978-1-4244-3521-0/09/$25.00 ©2009 IEEE

Step4. Every missing value a_p has a candidate value set S= {A, T, G, C}, which value to choose depends on the computation result of the possibility for each base. Equation (11) illustrates the computing method of possibility for A.

$$\Pr o(a_p = A) = \sum_{i=1, b_i^{(p)} = A,T}^{k} \beta * w_i \tag{11}$$

In equation (11), β can be compute with equation (12):

$$\beta = \begin{cases} 0, & b_i^{(p)} = T, d(a,b_i) > 0 \mid b_i^{(p)} = A, d(a,b_i) < 0 \\ 1, & b_i^{(p)} = A, d(a,b_i) > 0 \mid b_i^{(p)} = T, d(a,b_i) < 0 \end{cases} \tag{12}$$

Other possibility for candidate value like T, C and G can also be obtained referring to equation (11). After all the possibilities is worked out, the candidate value with maximum possibility can be assigned to a_p as estimated value, as showed in equation (13).

$$a_p = \{ s_i \mid \begin{matrix} s_i \in S, \forall s_j \in S \\ \Pr o(a_p = s_i) \geq \Pr o(a_p = s_j) \end{matrix} \} \tag{13}$$

IV. EXPERIMENTS AND RESULTS

A series of comparison will be displayed to explain the performance of our work. All the data is from genbank which can be downloaded from http://www.ncbi.nlm.nih.gov.

All of the genes in downloaded dataset are complete genes, for experimental purpose, a copy of the original data was made, and then, about 1%~5% of genes was made partly missing, thus, the missing rate denotes the percentage of incomplete gene in dataset. To evaluate the quality of estimation, accuracy rate is adopted which can be computed using equation (14).

$$Acr = \frac{N_{correct}}{N_{missing}} * 100 \tag{14}$$

With missing rate grows from 1.5% to 5%, the performance of traditional KNN and improved KNN is showed in figure 1; the number of neighbors is 10.

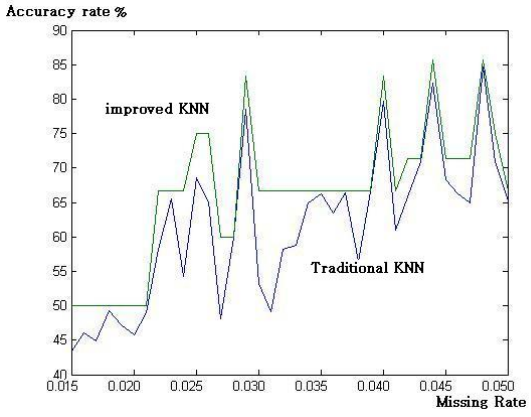

Figure 1. The performance of traditional KNN and improved KNN with missing rate from 1.5% to 5%.

It can be seen from figure 1, the performance of improved KNN outperforms traditional KNN in both accuracy rate and stability. The performance of both algorithms was better when missing rate was around 4% and 4.5%. Another comparison is illustrated in Figure 2 with the neighbor number grows from 10 to 50 and missing rate of 2.3%.

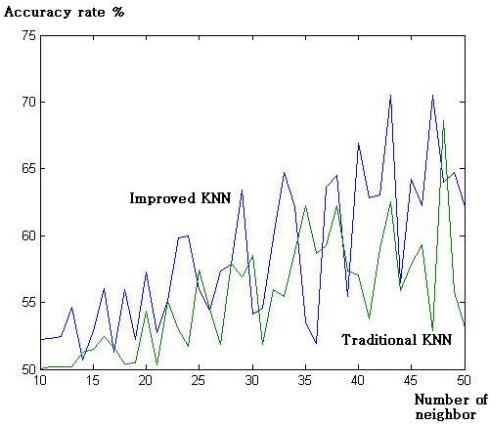

Figure 2. The performance of tr aditional KNN and improved KNN with 2.3% missing rate

As showed in figure 2, improved KNN performs better than traditional KNN most of the time, but both algorithms fail to give a stable result, another comparison is showed in figure 3 under the condition that the number of neighbors grows from 10 to 50 and missing rate to be 4.5%.

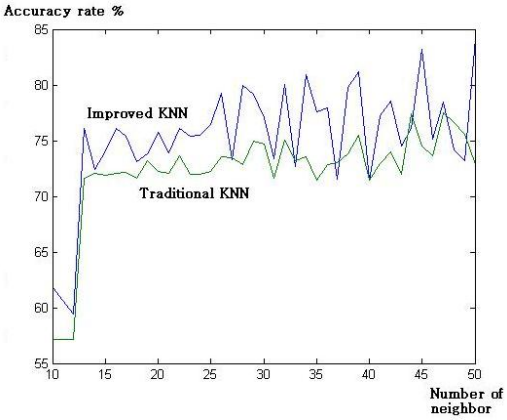

Figure 3. The performance of tr aditional KNN and improved KNN with 4.5% missing rate

In general, the performance of both algorithms is acceptable, but improved KNN works better than traditional KNN in accuracy, the accuracy rate generally grows as the number of neighbors grows, but the influence of missing rate of dataset is not that obvious under missing value less than 5%.

V. FUTURE WORK

KNN is one of the common classification algorithms in data mining, it has a wide range of applications in many areas for its simplicity. It was characteristic for gene data to have large size and close relation between samples, the application of KNN in gene data can archive preferably effect. However, there are other forms of mutation in genes, including replacement and

insertion, these mutation can bring difficulty in estimating missing values in gene data, how to improve the performance of algorithm while reduce the influence by other kinds of mutation will be another problem worth consideration.

REFERENCES

[1] Dudani, Sahibsingh A., "The Distance-Weighted k-Nearest-Neighbor Rule," IEEE Transactions on Systems, Man and Cybernetics, Vol. 6, pp. 325-327, April 1976.

[2] Pedro J. García-Laencinaa, José-Luis Sancho-Gómeza, Aníbal R. Figueiras-Vidalb and Michel Verleysen, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation," Neurocomputing, Vol. 72, pp. 1483-1493, March 2009.

[3] ZHOU Qinghua, CHEN Jun et al., "A study on the allelic deletion and mutation of FHIT gene in human non-small cell lung cancer, Chinese Journal of Lung Cancer. China," Vol. 4, pp. 10-14, February 2001.

[4] WANG Ye, GUAN Jinping, GAO Wei, et al. "Fragile histidine triad gene deletion in breast cancer and its correlation with human papillomavirus infection," Chinese Journal of General Surgery. China, Vol. 18, pp. 168-170, March 2003.

[5] LIU Peng, LEI Lei, ZHANG Xue-Feng. "A Comparision Study of Missing Value Processing Methods," Computer Science. China, Vol. 10, pp. 155-156, October 2004.

[6] Carol Featherstone1 and Stephen P Jackson. "DNA double-strand break repair," Current Biology, Vol. 9, pp. R759-61, October 1999.

[7] SALIMA HACEIN-BEY-ABINA, FRANÇOISE LE DEIST, FRÉDÉRIQUE CARLIER, et al. SUSTAINED CORRECTION OF X-LINKED SEVERE COMBINED IMMUNODEFICIENCY BY EX VIVO GENE THERAPY, The New England Journal of Medicine, Vol. 346, pp. 1185-1193, April 2002.

 978-1-4244-3521-0/09/$25.00 ©2009 IEEE