K-Ranked Covariance Based Missing Values Estimation for Microarray Data Classification

Muhammad Shoaib B. Sehgal, Iqbal Gondal and Laurence Dooley
GSCIT, Monash University, VIC 3842, Australia
{Shoaib.Sehgal, Iqbal.Gondal, Laurence.Dooley}@infotech.monash.edu

Abstract

Microarray data often contains multiple missing genetic expression values that degrade the performance of statistical and machine learning algorithms. This paper presents a K ranked diagonal covariance-based missing value estimation algorithm (KRCOV) that has demonstrated significantly superior performance compared to the more commonly used K-nearest neighbour (KNN) imputation algorithm when it is applied to estimate missing values of BRCA1, BRCA2 and Sporadic genetic mutation samples present in ovarian cancer. Experimental results confirm KRCOV outperformed both KNN and zero imputation techniques in terms of their classification accuracies when used to impute randomly missing values from 1% to 5%. The classifier used for this purpose was the Generalized Regression Neural Network. The paper also provides a hypothesis for why KRCOV performs better than KNN not only for bioinformatics data but also for other data types having strong correlated values.

1. Introduction

Microarray data is used for the simultaneous study of multiple genes under different conditions [15]. The application of DNA microarrays includes the study of human tumours [10] and Yeast sporulation (Chu 1998). Various machine learning algorithms are used in the molecular classification of DNA microarray data. Golub (1999), Ship (2002), Pomery (2002), Bhattacharjee (2001) and Ramaswamy (2001) all provide a wide range of examples of machine learning algorithms have been applied to leukemia, lymphoma, brain cancer, lung cancer and multiple primary tumour classification.

Despite its wide usage however, 90% of the time microarrays unfortunately generate missing values in the data [5]. Missing values occur for various reasons including image artifacts, slide scratches, insufficient resolution and hybridization errors (15, 5], which all leads to a degradation in the performance of many algorithms such as Principal Component Analysis, Singular Value Decomposition [11], Clustering, Classification and other statistical algorithms [1, 8]. This problem can be handled in several ways; one is to repeat the experiment which is not feasible for economic reasons, while an alternative is to remove the samples containing missing values, but this may often be inappropriate because in bioinformatics we have a limited number of samples. So the best solution is to estimate the missing values. The most common methods used so far are zero impute (replace missing values with zero) [1], row averages (replace missing value by its row average) and row median (replace missing value by a row median). However, Troyanskaya [14] demonstrated that these methods do not exploit the correlation between data in estimating missing values, which could improve the performance of classifiers.

This paper presents a missing value estimation technique called *K*- Ranked Covariance based missing value estimation (KRCOV), which is based on the principle that the missing values of particular genes can be estimated using most correlated genes. We compared our proposed estimation technique with the popular imputation technique K- Nearest Neighbour (KNN).

The well known ovarian cancer microarray data by Amir [2] is used for comparative purposes. The motivation to use this data is that ovarian cancer is the fourth most common cause of cancer-related deaths in American women of all ages, as well as being the most prevalent cause of death from gynecologic malignancies (NCH 1991). A reliable test for the type of mutation detection will be a significant help for the immunity of the cancer. Mutations in BRCA1, BRCA2 and Sporadic (without BRCA1 and BRCA2 mutation) can lead to carcinogenesis through different molecular pathways so disease pathway mapping is very helpful for the treatment of this disease.

Tests were conducted by randomly removing 1-5% values from the BRCA1, BRCA2 and Sporadic mutation data (mutations present in ovarian cancer) [2]. The technique has been compared not only based on estimation errors but also on classification errors of the above described mutations. The classifier used for this purpose was the Generalized Regression Neural Network (GRNN) because it demonstrates a high capability to classify ovarian cancer genetic mutations [11]. In this paper, we will also demonstrate a new estimation method,

Proceedings of the Fourth International Conference on Hybrid Intelligent Systems (HIS'04)
0-7695-2291-2/04 $ 20.00 IEEE

which is quantitatively superior to the KNN imputation method.

The rest of this paper is organized as follows. In Section 2 imputation and classification techniques are presented. Section 3 presents the novel KRCOV technique and theoretical basis that why KRCOV performs better than KNN, while experimental methodology is described in Section 4. Section 5 discusses missing value estimation results and their impact on the classification accuracy. Some conclusions are given in Section 6.

2. Applied Imputation and Classification Methods

This Section presents theoretical basis for the classification and missing value imputation methods that are to be compared with KRCOV..

2.1. GRNN Classification for Ovarian Cancer

Generalized Regression Neural Networks (GRNN) are paradigms of the Radial Basis Functions (RBF) used in functional approximation [13, 14]. To apply GRNN to classification, an input vector x (BRCA1, BRCA2 or Sporadic genetic data) is formed and weight vectors W are calculated using (2). The output (BRCA1, BRCA2 or Sporadic) *y(x)* is the weighted average of the target values t_i of training cases x_i close to a given input case *x*, as given below:-

$$y(x) = \frac{\sum_{i=1}^{n} t_i W_i}{\sum_{i=1}^{n} W_i} \qquad (1)$$

$$\text{where } W_i = \exp\left[\frac{-\|x - x_i\|^2}{2h^2}\right] \qquad (2)$$

The only weights that are needed to be tuned are the smoothing parameters *h* of the RBF units in (2), which are defined using a simple grid search method [17].

The GRNN is trained such that the Target Class *M* is selected from the set of classes required to be identified. The system is trained on a subset of samples T_r labeling these as positive examples (target value 1). The subsets of data for remaining classes serve to provide negative examples to the system (target value 2).
The distance between the computed value *y(x)* and each value in the set of target values *T* is given by:-

$$T = \{1, 2\} \qquad (3)$$

The values 1 and 2 correspond to the training class and all other classes respectively. The class corresponding to the target value with least square distance is chosen.

2.2. K- Nearest Neighbour (KNN) Estimation

The K-Nearest Neighbour (KNN) method selects genes with expression values similar to those genes of interest to impute missing values [15]. In order to estimate the missing value Y_{IJ}, of gene *I* and experiment *J,* *k* genes are selected whose expression vectors are similar to YIJ. The similarity measure between the vectors Y_1 and Y_2 is defined by the Euclidian distance reciprocal (see (4)) over all observed components in experiment *J*.

$$\psi = 1 / \| Y_1 - Y_2 \| \qquad (4)$$

The missing value is then estimated as the weighted average of the corresponding entries in the selected *k* expression vectors using (5).

$$\hat{Y}_{IJ} = \sum_{i=1}^{k} W_i . X_i \qquad (5)$$

$$W_i = \frac{1}{\psi_i \times \Delta} \qquad (6)$$

Where $\Delta = \sum_{i=1}^{k} \psi_i$ and ψ is the Euclidean distance. (6) shows the contribution of each gene is weighted by a similarity of its expression to that gene *I*.

The KNN based imputation method has no theoretical criteria for selecting the best *k*-value and distance function. Both k-value and distance function have to be determined empirically. Choosing a small *k* value produces a poorer classifier performance after imputation due to overemphasis of a few dominant instances in estimating the missing values. Conversely, a large neighborhood may include instances that are significantly different from those containing missing values impacting upon the estimation accuracy and hence classifier performance. Empirical results show that for small datasets *k* = 10 should be used [1]. Troyanskaya [14] demonstrated that Euclidean distance function performs best for KNN as this measure is sensitive to outliers that may be present in the microarray data, though our investigations revealed that log-transforming the data reduced their effect on gene similarity determination [15] These empirical selections of the *k*- value and distance function make the model difficult to use and less reliable.

3. Covariance Based Missing Value Estimation

The new *k*-ranked covariance-based method presented in this paper estimates missing values using expression values of genes having covariate genetic expression profiles of the target gene. To estimate missing value Y_{EG} of gene expression matrix *M* of experiment *E* and gene *G*, the absolute diagonal covariance *CoV* [10] of expression values of *G* is calculated with remaining genes having no missing value. Using the absolute covariance means that the higher the absolute covariance values, the more the genes are related. Those genes with missing values in E are ignored in the estimation. The *CoV* values are then ranked, and expression values of K, which are the most correlated genes $\hat{C}$, are selected from *E*. These expression values of *K* covariate genes are used to estimate Y_{EG} using (12), with *Wi* given by (8).

$$Y_{EG} = \sum_{i=1}^{k} W_i \hat{C}_i \qquad (7)$$

$$W_i = \frac{1}{(\hat{C}_i \times \gamma) + \zeta} \qquad (8)$$

Where ς is set to 10-2 to avoid divide by zero.

$$\gamma = \sum_{i=1}^{k} \hat{C}_i \qquad (9)$$

3.1. Reasons why KRCOV Performs Better than KNN

There are two fundamental reasons why KRCOV gives better estimation than KCOV.
Firstly KRCOV uses both negative and positive correlation values while KNN due to Gaussian distance function, only searches for positive covariate values. In estimating missing gene values it therefore ignores those genes which are inversely proportional to each other.

Lemma 1: KNN only considers positive correlations.
Proof: If there are 2 sets α and β which are inversely proportional to each other, then the distance *d* between α and β will be larger in those sets which are directly proportional to each other. Several distance functions are used for KNN and the most common is Gaussian which is given by:-

$$d = \|\alpha - \beta\| \qquad (10)$$

Which always results in higher value of d when α is inversely proportional to β.

Lemma 2: KRCOV considers both positive and negative correlation values.
Proof: Assume two sets α and β which are inversely proportional, so $cov < 0 \forall \alpha, \beta$ where

$$\text{cov} = \frac{1}{(n-1)} \sum_{i=1}^{k} (\alpha_i - \bar{\alpha})(\beta_i - \bar{\beta}) \qquad (11)$$

From (11) it is clear that if a high correlation exists between the gene values (either directly proportional and positively correlated values or inversely proportional and negatively correlated values) there will be a higher absolute *cov* value.
The second reason for the better performance of KRCOV is that it exploits estimated *a priori* values in the estimation of current missing values.

Lemma 3: KRCOV gives better estimation of missing values in case of transitive gene dependency (Gene $A \rightarrow B \rightarrow C$) than KNN.

Proof: Assume in an experiment *E* gene G_{a1} is correlated with S_1. as:-

$$G_{a1} \rightarrow S_1 \text{ such that } S_1 = \{G_{b1}, G_{b2} ... G_{bn}\} \qquad (12)$$

Similarly gene Gb1 is correlated with S2. as:-

$$G_{b1} \rightarrow S_2 \text{ such that } S_2 = \{G_{c1}, G_{c2} ... G_{cn}\} \qquad (13)$$

If the values of both G_{a1} and G_{b1} are missing then we can predict expression value of G_{b1} using set S_2 and finally predict the value of Ga1 more accurately using S1 including G_b rather than ignoring it. However, KNN based estimation does not consider estimated values in predicting future missing values.

4. Methodology

Ovarian cancer microarray data [2] was used in our experiments. The data set contains 18, 16 and 27 samples of BRCA1 mutations, BRCA2 and sporadic mutations (neither BRCA1 nor BRCA2) respectively. Each data sample contains logarithmic microarray data of 6445 genes. The missing value estimation techniques were tested using the following two approaches.

4.1. Imputation Error Estimation

This involves randomly removing values from the data and then computing the estimation error. For test purposes, between 1% and 5% of the values were removed from each the BRCA1, BRCA2 and Sporadic dataset samples and the Root Mean Square (RMS) errors ξ of imputation computed as :-

$$\xi = \frac{RMS(M - M_{est})}{RMS(M)} \qquad (14)$$

where M is the original data matrix and M_{est} is the estimated matrix using either KNN or KRCOV. The advantage of using (11) for error estimation is that ξ=1 for zero imputation [5].

4.2. Classification Error Estimation

An alternative way to test imputation method is by randomly removing values from the data and testing the impact on decision making techniques like classification and clustering. Thus between 1% and 5% data values were randomly removed and tested using GRNN by zero value imputation, KNN and KRCOV. The validation results were generated using k-fold cross validation. To correctly identify the classification accuracy, the data was divided evenly into k folds and the system processed for k-iterations. For each k^{th} experiment, k-1 folds were used for training and just one for testing such that the selection probability P_v of each fold to become a part of validation data for a particular iteration is:

$$P_v = \frac{\eta}{N \times L} \tag{15}$$

while the probability P_t of the remaining subset being selected as training data for a particular iteration is:

$$P_t = \frac{(k-1) \times \eta}{N \times L} \tag{16}$$

where N = total data items per class, k = number of folds, L = number of classes and η = number of samples in each subset.

After k iterations all subsets will have been part of the validation set, so the overall probability of the data as test data is unity, thereby giving the results a higher confidence level. The classification *accuracy* is given by:-

$$Accuracy = \frac{1}{k} \sum_{i=1}^{k} Acc_i \tag{17}$$

Where *Acc* is the intermediate accuracy after each iteration.

The motivation to use k-fold cross validation over the classical hold out or random resampling methods was that it uses data sets evenly both for training and testing, thereby giving better estimation of the classification rates.

5. Discussions of Results

As mentioned in Section 4, to test the new imputation technique, the *Imputation Error* and C*lassification Error Estimation* techniques were used. This section discusses the results of these experiments.

5.1. Imputation Error Estimation

Different values of K were tested for both KNN and KRCOV, with K =10 exhibiting the best results. The plots in Figure 1 show the RMS error for the estimation of randomly missing values from 1% to 5% for BRCA1 samples which confirms that KRCOV performed significantly better than KNN. It is also clear from the graphs that KRCOV exhibits far better robustness for higher missing values compared with KNN.
Similarly, Figure 2 depicts the RMS Error for the estimation of missing values for the BRCA2 data set, with again the results endorsing the improved performance of KRCOV over KNN especially at higher percentages. A similar observation is also apparent in Figure 3 for the Sporadic data set.

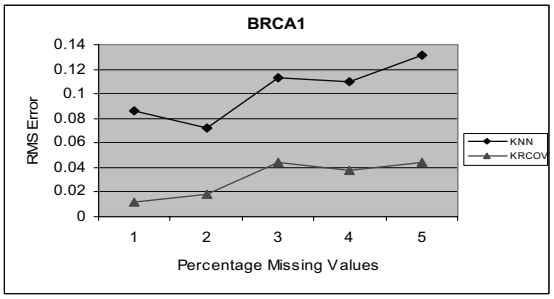

Figure 1: Estimation RMS error for BRCA1 with *k*=10

The reason of these improved results is that if there is a relationship between gene expression values then they can be predicted by calculating the degree of variation with respect to each other. KRCOV determines relationship between gene expression values in a better way than KNN due to the reasons provided in Section 3.1.

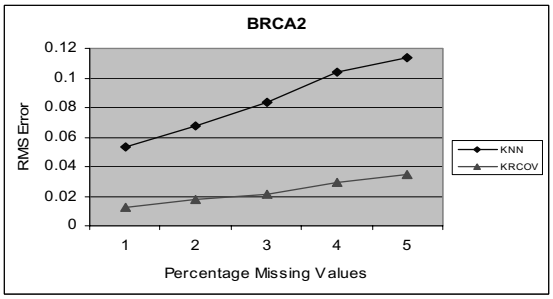

Figure 2: Estimation RMS error for BRCA2 with *k*=10

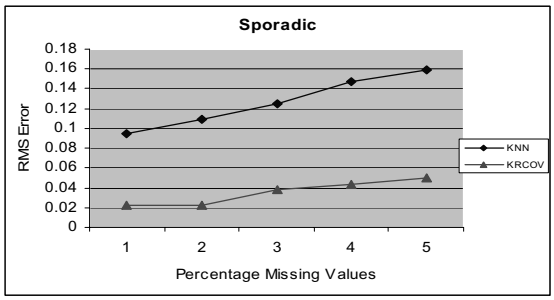

Figure 3: Estimation RMS error for Sporadic with *k*=10.

5.2. Classification Error Estimation

Missing values inevitably affect classification accuracy, yet many classifiers simply use zero Imputation [8]. However, our cross validation results show that with proper estimation of missing values classification accuracy can be significantly improved [11].

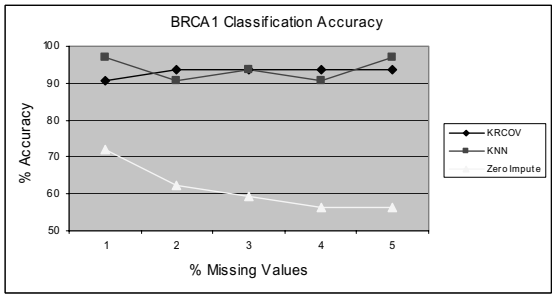

Figure 4: Classification Accuracies of BRCA1 Mutation

We tested the classification error of GRNN by introducing various missing values percentage by randomly removing values. The imputation method used were zero impute, KNN impute and the new KRCOV impute. Different GRNN trained in one-verses-all method were tested for the classification of BRCA1, BRCA2 and Sporadic genetic mutations by using the above imputation methods. Figure 4 shows that KRCOV and KNN both performed much better than zero imputation due to the accurate estimation of missing values. Results in Figure 5 demonstrate that KRCOV has outperformed KNN and zero imputation. Similarly, Figure 6 shows the same observation however zero impute has performed better than KRCOV when Sporadic samples had 2 percent missing values. The reason for this is that if the data between classes is more separable and thus easier to classify, i.e. the zero values included actually improved separability. In reality however, zero imputation does not improve separability most of the time. For example, if a particular gene has missing values, for both classes to be classified, zero imputation will result in the same value (i.e. zero). This means the gene has same value for both classes despite some genes being more significant than others. For this reason it is always better to have accurate estimation of missing values.

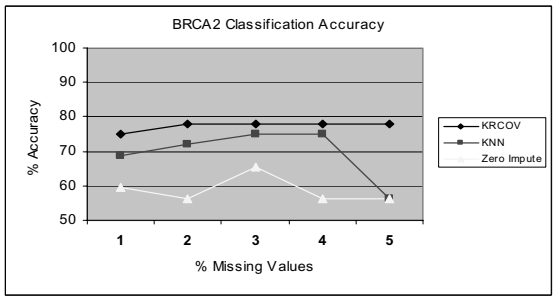

Figure 5: Classification Accuracies of BRCA2 Mutation

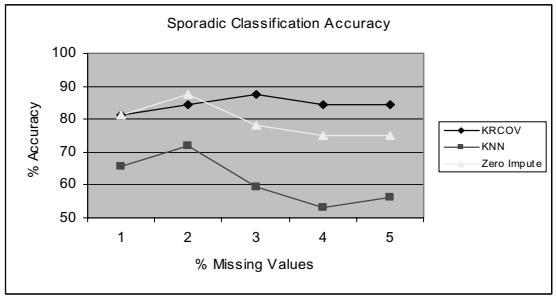

Figure 6: Classification Accuracies of Sporadic Mutation

6. Conclusions

This paper has presented a novel diagonal covariance-based (KRCOV) algorithm for missing value estimation, with its performance analyzed using ovarian cancer microarray data. The results show that KRCOV compared very favorably with K-nearest neighbour (KNN) and zero impute methods, both in terms of normalized RMS error rates and classification accuracy. This improvement is significant as it has been previously demonstrated that KNN performs better than other techniques including row average and zero impute. Also, while KRCOV performed better for ovarian cancer microarray data it is likely to performed significantly better for estimation for other bioinformatics data because of its ability to find the correlations between rows in the data.

7. References

[1] Acuna, E. and Rodriguez, C. (2004), "The treatment of missing values and its effect in the classifier accuracy", Classification, Clustering and Data Mining Applications. Springer-Verlag Berlin-Heidelberg, pp. 639-648.

[2] Amir A. J., Yee C. J., Sotiriou C., Brantley K. R., Boyd J., Liu E. T. (2002), "Gene expression profiles of brca1-linked, brca2-linked, and sporadic ovarian cancers", Journal of the National Cancer Institute, vol. 94 (13).

[3] Bhattacharjee A., Richards W. G., Staunton J., Li C., Monti S., Vasa P., Ladd C., Beheshti J., Bueno R., Gillette M., Loda M., Weber G., Mark E. F., Lander E. S., Wong W., Johnson B. E., Golub T. R., Sugarbaker D. J. & Meyerson M. (2001), "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses". Proc. Natl. Acad. Sci. USA 98: pp.13790–13795.

[4] "National center for health statistics: vital statistics of the United States" (1991), National Center for Health.

[5] Ouyang M, Welsh WJ, Georgopoulos P (2004), "Gaussian mixture clustering and imputation of microarray data", Bioinformatics.

[5] Pomeroy S. L., Tamayo P., Gaasenbeek M., Sturla L. M., Angelo M., McLaughlin M. E., Kim J. Y., Goumnerova L. C., Black P. M., Lau C., Allen J. C., Zagzag D., Olson J., Curran T., Wetmore C., Biegel J. A., Poggio T., Mukherjee S., Rifkin R., Califano A., Stolovitzky G., Louis D. N., Mesirov J. P., Lander E. S. & Golub T. R. (2002), "Prediction of central nervous system embryonal tumour outcome based on gene expression". Nature, 415(24): pp. 436-442.

[6] Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D. & Lander E. S. (1999), "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring". Science, 286(5439):pp. 531-537.

[7] Gustavo B., Monard C.M. (2003), "An analysis of four missing data treatment methods for supervised learning", Applied Artificial Intelligence 17(5-6): pp. 519-533.

[8] Ramaswamy S., Tamayo P., Rifkin R., Mukherjee S., Yeang C. H., Angelo M., Ladd C., Reich M., Latulippe E., Mesirov J. P., Poggio T., Gerald W., Loda M., Lander E. S. & Golub T. R. (2001), "Multiclass cancer diagnosis using tumour gene expression signatures", Proc. Natl. Acad. Sci., USA, 98(26):pp. 15149-15154.

[9] Ruth M. Mickey, Olive Jean Dunn, Virginia A. Clark (2004), Applied Statistics: Analysis of Variance and Regression, 3rd ed., Wiley Series in Probability and Statistics.

[10] Shoaib M. B. S, I. Gondal, L.Doloey, (2004), "Statistical neural networks and support vector machine for the classification of genetic mutations in ovarian cancer", to be published in Proc. IEEE CIBCB 04.

[11] Shipp M. A, Ross K. N., Tamayo P., Weng A. P., Kutok J. L, Aguiar R. C., Gaasenbeek M., Angelo M., Reich M., Pinkus G. S., Ray T. S., Koval M. A., Last K. W., Norton A., Lister T. A., Mesirov J., Neuberg D. S., Lander E. S., Aster J. C. & Golub T. R. (2002), "Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning", Nat Med, 8(1):pp. 68-74.

[12] Specht D. F. (1991), "A generalized regression neural network", IEEE Trans. on Neural Networks, pp. 568-576.

[13] Timmerman D., Verrelst H., Bourne T. H., De Moor B., Collins W. P., Vergote I. & Vandewalle J. (1999), "Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses". Ultrasound Obstet Gynecol, 13: pp. 17- 25.

[14] Troyanskaya O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R. B. Altman (2001), "Missing value estimation methods for DNA microarrays", Bioinformatics, vol.17, 520—525.

[15] "World Health Statistics Annuals" (1992), World Health Organization, Geneva, Switzerland.

[16] Lomax, A., J. Virieux, P. Volant and C. Berge, (2000), "Probabilistic earthquake location in 3D and layered models: Introduction of a Metropolis-Gibbs method and comparison with linear locations, in Advances in Seismic Event Location" Thurber, C.H., and N. Rabinowitz (eds.), Kluwer, Amsterdam, pp. 101-134.