ISSN:0975-9646

Improvement in Estimation of Unidentified Value (Gene Expression Data) in Biotechnology

Baitali Nath, Bindu Agarwalla, Laxman Sahoo
School of Computer Engineering, KIIT University, Bhubanseswar-751024,Odisha, India

***Abstract*: In this era, DNA microarray technology is used combining with different data mining processes for extracting relevant knowledge from genes of organisms to discover the association between noble diseases and their correlated genes. However, this gene expression data frequently contains absent values which are to be dealt with to stop them from causing drastic affect in further analysis processes. To overcome the same, a number of missing-value recovery approaches are being introduced to serve the purpose. In this paper, a Clustering Approach of Collaborative Filtering is projected to estimate missing values more precisely than done by existing approaches. The Collaborative Filtering used in the process, which is primarily used in Recommender Systems, has been united with a basic clustering method based on Rough-Set Theory to impute a missing value.**

Keywords: Collaborative filtering, Clustering method, Gene expression data, missing values, imputation.

1. INTRODUCTION

In present years, DNA microarray Gene Expression data are being extensively used in numerous fields to resolve the relationship between noble diseases and their related genes. Some effective processes have been introduced that made it possible to monitor numerous gene expression levels of organisms concurrently under different conditions [4,5,6]. In biotechnology, DNA Microarray Gene Expression Data Analysis is used for pharmaceutical use, cancer classification, protein sequencing and also in classification of genes which help in diagnosing certain disorders. However, DNA Microarray Gene Expression Data frequently include misplaced values for reasons such as corrupted image[20], error born due to hybridization, dust infected and insufficient resolution of the source. Unfortunately, these lost values extensively affect Gene Expression Data Analysis results. Huge amount of information is mislaid when genes with missing values are ignored or directly omitted.

So, in order to deal with the unidentified values in the Gene Expression dataset, some Imputation Techniques have been invented to estimate the unavailable values before conducting the actual data analysis. The k-Nearest Neighbor (k-NN) Method[8], the local least-square(LLS) approach [9], the Gene Ontology k-Nearest Neighbor (GOKNN) method [11,13] are among the few mostly used techniques in the process. Some other imputation algorithm includes SVD impute method[8], the Bayesian approach[10], the Collateral Missing Value Imputation Approach[12] etc. Even though these approaches work well but they also have some limitations. *k*-NN Imputation Method predicts best on Non-Time Series Data or noisy Time Series Data, whereas SVD impute approach gives good performance on Time Series Data with low noise levels and with a strong Global Correlation Structure. GOKNN is seen to be most efficient when strong local correlation exists in the dataset and so on.

The collaborative Filtering (*CF*) Approach is widely used in Recommender Systems that make prediction about a user taking into account the preferences of other users. Two types of basic *CF* algorithm are in use. First, memory-based (user-based) *CF* Algorithms, which recommends depending on the preferences between an active user and his top-*k* Nearest Neighbors. Second, model-based *CF* algorithms that recommends based on the training provided by its training dataset.

In this paper, the collaborative filtering technique, based on rough set theory, is used on the gene expression dataset to predict any unavailable or missing value of a condition in the dataset. A k-means clustering method is being combined with the collaborative filtering to reduce the sample space of the available data which are used for predicting the missing value. The main advantage of this technique would be an attempt to reduce the number of calculations to be performed on the dataset and hence reducing the total time taken by the process.

2. MOTIVATION

Although there exists various methods for estimating the absent values in Gene Expression Data, there are always some more and better efficient methods to do the same. Collaborative Filtering, which is the most trending technique used in recommender system in the present era is now also being implemented in biological field's data mining and analysis as well[1,2]. The main motivation for this paper is the need to reduce the high computational time in the user-based Collaborative Filtering technique.

3. RELATED WORK

Some imputation techniques developed to estimate the unavailable values in the gene expression dataset before conducting the actual data analysis are discussed here. The k-nearest neighbor (k-NN) method[8] uses the weighted average of the top k similar genes' values to impute the missing value of the concerned gene. The Local Least-Square(LLS) Approach [9] selects the top k-Nearest Neighboring genes and then predicts the missing values using the Least Square Method. The Gene Ontology k-Nearest Neighbor (GOKNN) Method [17,13] imputes the missing gene expression value by calculating the semantic similarity between 2 genes from their gene ontology annotations. Some other imputation algorithm includes SVD impute method[8], the Bayesian approach[10], the collateral missing value imputation approach[12] etc. Even

though these approaches work well but they also have several limitations. *k*-NN Imputation Method predicts best on non-time series dataset or noisy time-series dataset, whereas SVD impute approach gives good performance on time series data which are less corrupted and with a strong global correlation structure. GOKNN is seen to be most efficient when strong local correlation exists in the data and so on. Many collaborative filtering algorithms are in use for recommender systems. The GroupLen System is a widely-used user-based *Collaborative Filtering* algorithm. But this type of *CF* algorithm has quite a high calculation time because it has to calculate the similarities between the active user and all other users in the dataset to make a prediction for the active user. To remove this problem, Cluster-Based *CF* Algorithms [15] have been proposed.

B. Sarwar et al.[16] proposed an item-based algorithms technique which analyze the user-item matrix for identifying relationships between different user items, and then using these relationships to provide users with valuable recommendation. This collaborative filtering approach for recommenders systems was found to provide better performance as well better quality than the already existing user-based algorithms.

Chuan et al. [11] solved the suggestion problem by combining user-based *CF* regression with item-based *CF* *f*iltering. Some other hybrid CF algorithm have also been proposed in recent years [19,22,23].

4. EXISTING SOLUTION

The Collaborative Filtering (*CF*) Algorithm is applied to the DNA Microarray Datasets, as the mentioned dataset is found to be similar to that used in Recommender Systems, to estimate missing values. The method used by Bo-Wen Wang et al.[1] by combining user-based *CF* based on rough-set theory to impute missing value in gene expression data is found to be quite efficient.

Rough set is the formal estimate of crisp set which is represented by a pair of set whose upper and lower value is the approximation of the original set. Thus, it not only considers the similarity between genes, but also is concerned with all the conditions of each gene. Thus a better prediction of missing value can be obtained.

Collaborative Filtering Based on Rough-Set Theory(CFBRST) Method is applied in two distinct phases, namely, preprocessing phase and prediction phase[1].

Preprocessing Stage: In order to use the rough-set-based method, the given dataset must be converted from numerical data to categorical values before filling all the missing data using user-based CF method, leaving only the value to be predicted.

Prediction Stage: In the prediction phase, the rough set based prediction uses where Pearson Correlation Coefficient on conditions for a given set of genes to determine a class attribute:

$$sim(i,tcond)= \frac{\sum_{j\in G}(y_{j,i}-\bar{y}_i)(y_{j,tcond}-\bar{y}_{tcond})}{\sqrt{\sum_{j\in G}(y_{j,i}-\bar{y}_i)^2}\sqrt{\sum_{j\in G}(y_{j,tcond}-\bar{y}_{tcond})^2}}$$

where $y_{j,i}$ denotes the value of gene *j* on condition *i*, is the average value of condition i; where $y_{j,tcond}$ denotes the value of gene *j* on target condition, is the average value of target condition and *G* is the set of given genes.

Then, the Elementary Set of the Class Attribute (or condition) is generated. The available Elementary Subset which contains the Active Gene (missing-value-containing gene) is selected from the Elementary Set of the Class Attribute. The target condition is then combined with the condition having second highest similarity with the target condition. Then, the algorithm partitions the gene dataset into an Elementary Set of the new combined condition, according to the condition values. If the subsets of the Elementary Set of new condition are all enclosed in elementary set of class attribute and the number of genes in the subsets exceeds the gene constraint, these subsets are marked as a potential Equivalence Class Set. The subsets, whose combined condition values are similar to the condition values of combined condition, are selected as the Equivalence Class Set from the potential equivalence class set. The algorithm runs iteratively until an equivalence subset is obtained.

Finally, the absent value is predicted. The predicted value is then transformed back to its numerical form by using the reverse formula which was used to transform from numerical to categorical data.

5. PROPOSED APPROACH

The common CF algorithms proceed in two phases. At the first phase, calculate the similarities between pairs of users and identify their neighbor. The recommendations are then generated for active user based on the aggregate of the ratings of the neighbors. The main focus for an excellent Imputing Missing-Value System is on predicting the absent values appropriately as well as using optimal time. The proposed approach is thus developed by combining k-means clustering and the existing CBRST method[1].

In this approach, we intend to cluster gene samples after the Pre-processing of dataset in Collaborative Filtering Based on Rough-Set Theory. Then we apply the Prediction Phase within the cluster, containing the active gene, to generate the missing value. For performing this clustering operation, k-means clustering is used because of its simplicity in implementation. Instead of using each of the gene samples in prediction phase as in CFBRST, we can work with only that cluster of gene which contains the active gene. The missing values of the gene are predicted according to CFBRST method within the selected cluster.

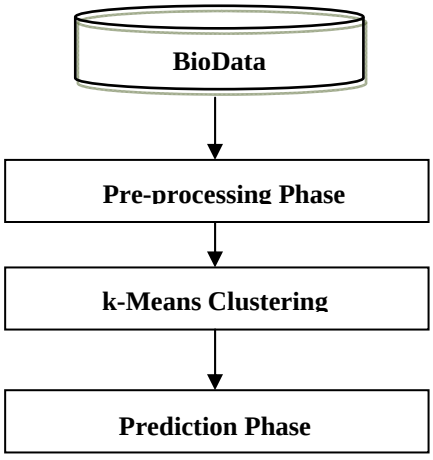

Fig. 1: Reference framework of the proposed approach

In this approach, first we intend to cluster the whole pre-processed dataset of gene samples. For performing this clustering operation, k-means clustering is used because of its simplicity in implementation. The following steps are to be followed:

1. The gene expression values which are farthest apart (using Euclidian distance measure), except the active gene sample, is chosen.
2. The remaining items are examined in order and allocated to the clusters to which they are closest.
3. Each time a new item is added to the cluster, the mean value of the cluster is modified.
4. Proper partitions of the clusters are obtained.
5. The active gene is checked against each cluster (Euclidian distance) by taking into consideration all the conditions except the missing one.
6. The active gene is then added to its nearest cluster.

Now, we will work with only those genes which belong to the cluster containing active gene and will thus eliminate the remaining genes in upcoming calculations. This combined approach may reduce the high-computational time taken for calculation of Pearson correlation coefficient. This may also reduce the number of iterations of the CFBRST algorithm to generate equivalence class set.

6. CONCLUSION

The k-means clustering algorithm of data mining is associated with the techniques of Collaborative Filtering Based on Rough Set-Theory in microarray gene expression data in an attempt to reduce total computational time. In future work, an algorithm is to be developed for the proposed idea and the result is to be compared with the existing CFBRST method to evaluate the efficiency of the new algorithm.

REFERENCES

[1] Bo -Wen Wang, Vincent S. Tseng, "Improving Missing-Value Estimation in Microarray Data with Collaborative Filtering Base on Rough-Set Theory", International Journal of Innovative Computing, Information and Control, 2012.

[2] M. C. Pham, Mr. Y. Cao, Mr. R. Klamma, Mr. M. Jarke, "A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis" , International Journal of Innovative Computing, Information and Control, 2011.

[3] Eytan Domany, " Cluster Analysis of Gene Expression Data," Journal of Statistical Physics, August 2002.

[4] J. L. DeRisi, V. R. Iyer and P. O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, vol.278, pp.680-686, 1997.

[5] S. Dudoit, Y. H. Yang, M. J. Callow and T. P. Speed, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, vol.12, pp.111-139,2002.

[6] D. J. Lockhart and E. A. Winzeler, Genome, gene expression and DNA arrays, *Nature*, vol.405, pp.827-836, 2000.

[7] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L.Loh, J. R. Downing, M. A. Caligiuri, C. D. BloomField and E. S. Lander, Molecular classification for cancer: Class discovery and class prediction by gene expression monitoring, *Science*, vol.286, pp.531-537, 1999.

[8] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. B. Altman, Missing value estimation methods for DNA microarray, *Bioinformatics*, vol.17, pp.520-525, 2001.

[9] H. Kim, G. H. Golub and H. Park, Missing value estimation for DNA microarray gene expression data: Local least squares imputation, *Bioinformatics*, vol.21, pp.187-198, 2005.

[10] S. Oba, M. A. Sato and I. Takemasa, A Bayesian missing value estimation method for gene expression profile data, *Bioinformatics*, vol.19, pp.2088-2096, 2003.

[11] C. Yu, J. Xu and X. Du, Recommendation algorithm combining the user-based classified regression and the item-based filtering, *Proc. of the 8th International Conference on Electronic Commerce*, pp.574-578, 2006.

[12] M. S. Sehgal, I. Gondal and L. S. Dooley, Collateral missing value imputation: A new robust missing value estimation algorithm for microarray data, *Bioinformatics*, vol.21, pp.2417-2423, 2005.

[13] J. Tuikkala, L. Elo, O. S. Nevalainen and T. Aittolallio, Improving missing value estimation in microarray data with gene ontology, *Bioinformatics*, vol.21, no.5, pp.566-572, 2006.

[14] S. Nagi, D.K. Bhattacharyya, J.K.Kalita, "Gene Expression Data Clustering Analysis:A Survey", Emerging Trends and Applications in Computer Science (NCETACS), October 2011.

[15] G. R. Xue, C. Lin, Q. Yang, W. Xi, H. J. Zeng, Y. Yu and Z. Chen, Scalable collaborative filtering using cluster-based smoothing, *Proc. of the 28th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.114-121, 2005.

[16] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, Item-based collaborative filtering recommendation algorithms, *Proc. of the 10th International World Wide Web Conference*, pp.285-295, 2001.

[17] P. W. Lord, R. D. Stevens, A. Brass and C. A. Goble, Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation, *Bioinformatics*, vol.19, pp.1275-1283, 2003.

[18] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining Concepts and Techniques, New Delhi: Elsevier, 2014

[19] C. Basu, H. Hirsh and W. Cohen, Recommendation as classification: Using social and content-based information in recommendation, *Proc. of the 15th National Conference on Artificial Intelligence*, pp.714-720, 1998.

[20] S. Oba, M. A. Sato and I. Takemasa, A Bayesian missing value estimation method for gene expression profile data, *Bioinformatics*, vol.19, pp.2088-2096, 2003.

[21] D. Jiang, C. Tang, A. Zhang, Cluster Analysis for Gene Expression Data:A Survey, IEEE Transactions On Knowledge And Data Engineering, vol. 16, no. 11, November 2004.

[22] Y. Blanco-Fernandez, J. J. Pazos-Arias, M. Lopez-Nores, A. Gil-Solla and M. Ramos-Cabrer, VATAAR: An improved solution for personalized TV based on semantic inference, *IEEE Trans- action on Consumer Electronics*, vol.52, pp.421-429, 2006.

[23] A. Popescul, L. Ungar, D. Pennock and S. Lawrence, Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments, *Proc. of the 17th Conference in Uncertainty in Artificial Intelligence*, pp.437-444, 2001