SPECTRAL RELEVANCE CODING IN MINING FOR GENOMIC APPLICATION

S.J.Saritha[1] and Prof.P.Govindarajulu[2]

[1]Dept of CSE, JNTUA CE Pulivendula,A.P , INDIA
sarithajntucep@gmail.com
[2]Dept of CSE, S.V.University, Tirupathi A.P., INDIA,
pgovindarajulu@yahoo.com

ABSTRACT

Most current gene detection systems are Bio-informatics based methods. Despite the number of Bio-informatics based gene detection algorithms applied to CEGMA (Core Eukaryotic Genes Mapping Approach) dataset, none of them have introduced a pre-model to increase the accuracy and time reduction in the different CEGMA datasets. This method enables us to significantly reduce the time consumption for gene detection and increases the accuracy in the different datasets without loss of Information. This method is based on feature based Principal Component Analysis (FPCA). It works by projecting data elements onto a feature space, which is actually a vector space that spans the significant variations among known data elements.

KEYWORDS

Gene detection system, PCA, KPCA, Spectral simulation

1. INTRODUCTION

Communication networks make physical distances worthless. People can communicate with each other through the networks without any restriction of the real distance. While we treasure the ease of being connected, it is also recognized that a gene users from one place can cause severe damages to wide areas. Generally a gene is defined as "any set of actions that attempt to compromise the integrity, confidentiality or availability of information resources." The identification of such a set of malicious actions is called gene detection problem. The Gene detection systems are an integral package in any well configured and managed computer system or network. Generally Gene detection systems may be some software or hardware systems that monitor the different events occurring in the actual system and analyzing them for effective detection.

There are two major approaches in gene detection: anomaly detection and misuse detection. Misuse detection consists of first recording and representing the specific patterns of genes, then monitoring current applications for such patterns, and reporting the matches. There are several developed models in misuse gene detection [1] [2]. They differ in representation as well as the matching algorithms employed to detect such threat patterns. Anomaly detection, on the other hand, consists of building models from normal data and then detects variations from the normal model in the observed data. The main advantage with anomaly gene algorithms is that they can

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 01–07, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4901

detect new forms of genes, because these new genes will probably deviate from the normal original behavior of genes [3].

There are many Gene detection systems developed for gene detection. But most of them apply an algorithm directly [4, 5, 6] on the rough data obtained from traffic or other local or remote applications which increases the consumption time. The CEGMA gene detection datasets [7] are an example for these algorithms. To overcome the draw back of high time consumption, a method was proposed for gene detection based on the principal component analysis (PCA) [8]. This method Extracts the main components (repetitive components) of the incoming dataset and performs the gene detection only for those components. However this method reduces the time consumption but reduces the accuracy. To overcome this drawback another method is proposed named as advanced PCA. This method accomplishes with the clusters of incoming dataset based upon their header information. Though this method increases the accuracy and reduces the time consumption but there is possibility to alter the incoming bio informatics at switching stages. Thus it can be considered as valid. To overcome this drawback there should be another parameter to analyze the incoming informatics. This paper proposes a method to overcome the drawback of previous method by introducing a new parameter called spectral simulation. This method performs the calculation of spectral nature of incoming gene data set if the header of the incoming data packet is not matched. The rest of this paper is organized as follows;

Section II gives the detailed description of PCA on the gene data set. Section III gives the cluster formation of incoming gene data based on the specific features. Proposed spectral simulation method is discussed in section IV. The results obtained are represented in section V and finally section VI concludes the paper.

2. PRINCIPAL COMPONENT ANALYSIS (PCA)

This section gives the complete illustration about the principal component analysis and also tells how to extract the important (repetitive) features of the complete incoming gene dataset. It is often used to reduce the dimension of dataset for easy exploration. It is concerned with explaining the variance-covariance structure of a set of variables through a few new variables which are functions of the original variables. Principal components are particular linear combinations of the p random variables. Let X_1, X_2, …, X_p be the P random variables representing p gene datasets with three important properties: (1) the principal components are uncorrelated, (2) the first principal component has the highest variance, the second principal component has the second highest variance, and so on, and (3) the total variation in all the principal components combined is equal to the total variation in the original variables X_1, X_2, …, X_p. They are easily obtained from an Eigen analysis of the covariance matrix or the correlation matrix of X_1, X_2, …, X_p [9].

Principal components from the covariance matrix and the correlation matrix are usually not the same. In addition, they are not simple functions of the others. When some variables are in a much bigger magnitude than others, they will receive heavy weights in the leading principal components. For this reason, if the variables are measured on scales with widely different ranges or if the units of measurement are not commensurate, it is better to perform PCA on the correlation matrix.

Let **R** be a p x p sample correlation matrix computed from n observations on each of p gene datasets X_1, X_2, …, X_p. If (λ_1, $\mathbf{e}_1$), (λ_2, $\mathbf{e}_2$), …, (λ_p, $\mathbf{e}_p$) are the p eigen value-eigenvector pairs of **R**, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, then the i^{th} sample principal component of an observation vector x=(x1,x2,…..xp) is

$$y_i = \mathbf{e}_i'\mathbf{z} = e_{i1}z_1 + e_{i2}z_2 + \ldots + e_{ip}z_p, \quad i = 1,2,\ldots,p$$ Where

$\mathbf{e}_i = (e_{i1}, e_{i2}, \ldots, e_{ip})'$ is the i^{th} eigenvector

And

$\mathbf{z} = (z_1, z_2, \ldots, z_p)'$ is the vector of standardized observations defined as

$$z_k = \frac{x_k - \overline{x}_k}{\sqrt{s_{kk}}}, \quad k = 1,2,\ldots,p$$

Where k_x and s_{kk} are the sample mean and the sample variance of the variable X_k.

The i^{th} principal component has sample variance λ_i and the sample covariance of any pair of principal components is 0. In addition, the total sample variance in all the principal components is the total sample variance in all standardized variables $Z_1, Z_2, \ldots, Z_p$, i.e., $\lambda_1 + \lambda_2 + \ldots + \lambda_p = p$

This means that all of the variation in the original dataset is accounted by the principal components. But this method allows only repetitive components to classify with the incoming data set at testing. This is very effective in reducing the computation time by decreasing the total size where as the main draw back of this method, it is not able to give accuracy because when there is data set testing which is not a repetitive one. To overcome this problem a cluster based PCA is proposed and also discussed briefly in next section.

3. K-PCA

This section provides the information about the PCA based on kernel (clusters) features. Like in PCA, the overall idea is to perform a transformation that will maximize the variance of the captured variables while minimizing the overall covariance between those variables. Using the kernel trick, the covariance matrix is substituted by the Kernel matrix and the analysis is carried analogously in feature space. An Eigen value decomposition is performed and the eigenvectors are sorted in ascending order of Eigen values, so those vectors may form a basis in feature space that explain most of the variance in the data on its first dimensions.

However, because the principal components are in feature space, we will not be directly performing dimensionality reduction. Suppose that the number of observations **m** exceeds the input dimensionality **n**. In linear PCA, we can find at most **n** nonzero Eigen values. On the other hand, using Kernel PCA we can find up to **m** nonzero Eigen values because we will be operating on an ***mxm*** kernel matrix [10]. When the external features of all variables are matched with the features of variables present in database the gene is said to be detected, otherwise the variables are allowed for further process. Though this method increases the accuracy and reduces the time consumption but there is possibility to alter the incoming bio informatics of the gene dataset at various switching stages. Thus it can be considered as valid gene. To overcome this drawback there should be another parameter to analyze the incoming informatics. The next section gives the information about the spectral properties of incoming gene dataset which are allowed to further process.

4. PROPOSED METHOD

This method overcomes the above mentioned problem by extracting the spectral features of the incoming gene dataset. This method allows the comparison of spectral features of incoming dataset along with the normal features. In this method the internal features of the incoming gene dataset are also going to be compared with the features of dataset in the database. Then only they are going to allow for further process. Before this the complete spectral features of the gene dataset are have to be evaluated. For this purpose the complete incoming dataset is going to be represented in Binary format (1's and 0's).After this each and every gene is represented with a bit vector. This paper assumes the spectral characteristics of a gene data set as,

1. No of switching states out of all bits. I.e. how much number of times the bits changed their state out of all bits.
2. The symmetry property.
3. The transition time taken from one bit to next bit.

This spectral property plays a vital role in this paper. Based upon these spectral properties the incoming dataset is going to be tested and allowed for further process The data set present in the personal computer is divided into clusters based upon their headers as shown below.

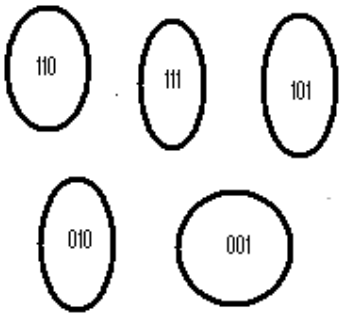
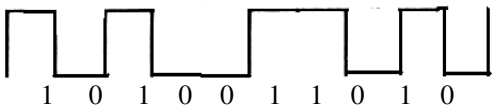


Figure1: clusters of data base

The incoming data set is compared with these clusters. If the header of a incoming data set is matched with any one of the clusters header it is detected as valid gene. This is processed out in previous approach.

In this approach first the present dataset is divided into clusters and also their spectral characteristics are calculated as follows:

Let a gene is represented with the bit vector shown below



The total number of bits=10
The total number switching states=7
The total switching ratio =7/10=.7

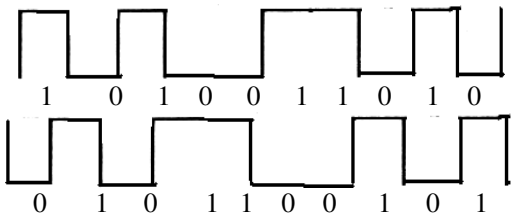


Figure2: spectral properties of dataset

Like this the total switching ratio is calculated for each and every data packet and kept in a cluster. When the incoming data set is said to matched with the header information of any cluster the spectral properties of that incoming dataset is also going to compared with spectral properties of the dataset. If they are matched then the incoming data set is said to be gene otherwise it is allowed for further process. The data is going to be switched by many steps during transmission. So there is possibility to change the header information intentionally and also non-intentionally. Non-intentionally means automatic change of header during transmission. There may be a possibility to change the header information by hackers also. This is referred to as intentional change. So the comparison of spectral properties of incoming dataset increases the accuracy as well as reduces the time consumption. The results discussed below gives the graphical information about this proposed method.

5. RESULTS

This section gives the illustration about the performance evaluation of the proposed method.

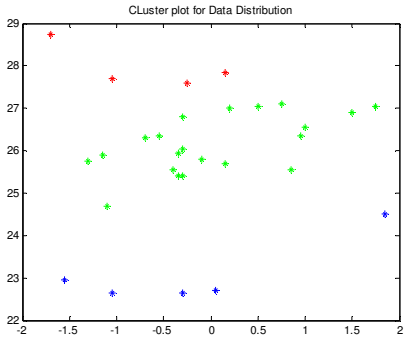


Figure 3: data scattering plot for relevance gene sequence

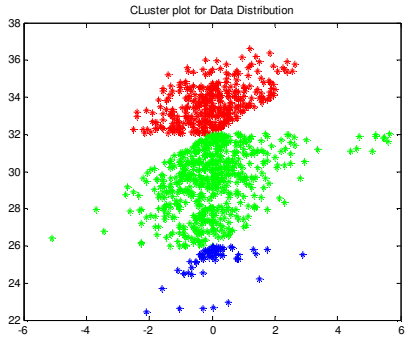


Figure 4: relevance clustering of genomic information in k class

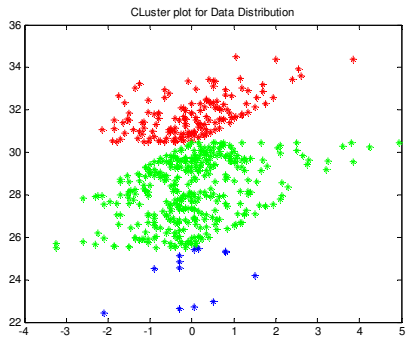


Figure5: relevance clustering of genomic data set for k-class after spectral mapping

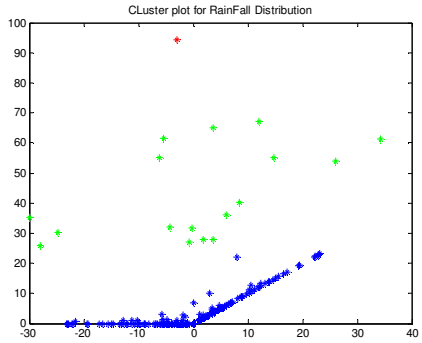


Figure6: cluster relevancy for a single class observation in 3–set data

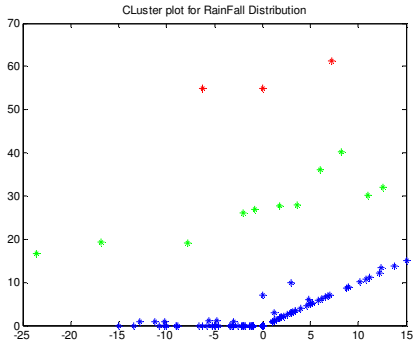


Figure 7: relevance plot of 1-class among 3-class Clustered data at spectral mapping

6. CONCLUSION

To improve the operation of data mining in this paper a spectral based doing approach is proposed. The proposed approach observes the variation in sequence pattern is developed and in similarity to a spectral correlation is observed. Pattern having sequence of similar spectral information is defined in bit pattern and a similar code is applied for representation to the existing coding pattern. For the test of such approach extended format of PCA called Kernel- PCA (K-PCA) is used. From the obtained observations it is observed that a improvement in processing efficiency with respect to Process time and recall efficiency is observed.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

REFERENCES

[1] K. Ilgun. Ustat, a real time gene detection system for UNIX. In IEEE Symposium on Security and Privacy, pages 16–28, Oakland, CA, May 1993.

[2] S. Kumar and E. Spafford. A pattern matching model for misuse gene detection. In Proceedings of the 17th National Computer security Conference, pages 11–21, 1994.

[3] D. Denning. An Gene Detection Model. IEEE Transactions on Software Engineering, 13(2):222–232, 1987.

[4] R. Agrawal and M. V. Joshi. PNrule: A New Framework for Learning Classifier Models in Data Mining A Case-Study in Network Gene detection. Technical Report RC-21719, IBM Research Division, 2000.
[5] I. Levin. CEGMA-99 Classifier Learning Contest LLSoft's Results Overview. SIGCEGMA Explorations. ACM SIGCEGMA, 1:67–71, 2000.
[6] B. Pfahringer. Winning the CEGMA Classification Cup: Bagged Boosting. SIGCEGMA Explorations.ACM SIGCEGMA, 1:65–66, 2000.
[7] CEGMA Cup 99 Gene Detection Datasets. Available at: http://CEGMA.ics.uci.edu/databases/CEGMAcup99/CEGMAcup99.html, 1999.
[8] I. T. Joliffe. Principal Component Analysis. Springer Verlag, New York, NY, third edition, July 2002.
[9] I.T.Jolliffe,"principal component analysis", Ed.,springer-verlag, NY, 2002
[10] FASEL, Ian. Scholkopf, Smola and Muller: KernelPCA. Available in: http://cseweb.ucsd.edu/classes/fa01/cse291/kernelPCA_article.pdf
[11] Lee, S.hyun. & Kim Mi Na, (2008) "This is my paper", ABC Transactions on ECE, Vol. 10, No. 5, pp120-122.
[12] Gizem, Aksahya & Ayese, Ozcan (2009) Coomunications & Networks, Network Books, ABC Publishers.

AUTHORS

Smt. S.Jessica Saritha is currently working as an Assistant Professor in Department of CSE JNTUA College of Engineering, Pulivendula,Andhra Pradesh India Her research interests are Data mining and Distributed databases

Prof. P. Govindarajulu is a rertired professor in department od Computer Science and Engineering Sri. Venkateswara University Tirupathi, Andhra Pradesh . HE worked at various portfolios in the university . His research interests of the Databases and Data mining