Volume 5, Issue 11, November 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Emerging Statistical Models for the Analysis of Genomic Data

S. Jessica Saritha
Dept of CSE, JNT University, Anantapur,
Andhra Pradesh, India

Prof. P. Govindarajulu
Department of CSE, Sri Venkateswara University,
Andhra Pradesh, India

Abstract: In this paper we review important emerging statistical models that have been recently developed and used for genomic data analysis. First, we summarize general background and some critical issues in genomic data mining. We then describe a novel concept of statistical significance, so-called false discovery rate, the rate of false positives among all positive findings, which has been suggested to control the error rate of numerous false positives in large screening biological data analysis. In the next section two recent statistical testing methods---significance analysis of microarray (SAM) and local pooled error (LPE) tests are introduced. We next introduce statistical modeling in genomic data analysis such as ANOVA and heterogeneous error modeling (HEM) approaches that have been suggested for analyzing microarray data obtained from multiple experimental and/or biological conditions.

Keywords: ANOVA, False discovery rate, Genomic data, Heterogeneous error model (HEM), Hierarchical clustering, Linear discriminate Analysis, Local pooled error (LPE) test, Logistic Regression discriminate analysis, Microarray Gene Chip™ gene expression, Misclassification penalized posterior (MiPP), Significance analysis of microarray (SAM), Supervised learning, Unsupervised learning

I. INTRODUCTION

There has been a great explosion of genomic data in recent years. This is due to the advances in various high-throughput biotechnologies such as RNA gene expression microarrays. These large genomic data sets are information-rich and often contain much more information than the researchers who generated the data may have anticipated. Such an enormous data volume enables new types of analyses, but also makes it difficult to answer research questions using traditional methods. Analysis of these massive genomic data has several unprecedented challenges:

II. CHALLENGES IN GENOMIC DATA ANALYSIS

Challenge 1: Multiple comparisons issue

Analysis of high-throughput genomic data requires handling an astronomical number of candidate targets, most of which are false positives [1, 2]. For example, a traditional statistical testing criterion of 5% significance level would result in 500 false positive genes, on average, from a 10K microarray study comparing two biological conditions for which there were no real biological differences in gene regulation. If there actually were a small number of, e.g. 100 genes that are differentially regulated, such *real* differentially expressed genes will be mixed with the above 500 false positives without any *a priori* information to discriminate the two groups of genes. Confidence on the 600 targets identified by such a statistical test is low, and further investigation of these candidates will have a poor yield. Simply tightening such a statistical criterion, e.g., a 1% or lower significance level, will result in a high false-negative error rate with failure to identify many important real biological targets. This kind of pitfall, the so-called *multiple comparisons issue* becomes even more serious when one tries to find novel biological mechanisms and biomarker prediction models that involve multiple interacting targets and genes, because the number of candidate pathways or interaction mechanisms grows exponentially. Thus, it is critical that data mining techniques effectively minimize both false positive and false negative error rates in these kinds of genome-wide investigations.

Challenge 2: High dimensional biological data

The second challenge is the high dimensional nature of biological data in many genomic studies [3]. In genomic data analysis, many gene targets are investigated simultaneously, yielding dramatically sparse data points in the corresponding high-dimensional data space. It is well known that mathematical and computational approaches often fail to capture such high dimensional phenomena accurately. For example, many search algorithms cannot freely move between local maxima in a high dimensional space. Furthermore, inference based on the combination of several lower dimensional observations may not provide a correct understanding of the real phenomenon in their joint, high-dimensional space. Consequently, unless appropriate statistical dimension reduction techniques are used to convert high dimensional data problems into lower dimensional ones, important variation and information in the biological data may be obscured.

Challenge 3: Small n and large p problem

The third challenge is the so-called “small n and large p” problem [2]. Desired performance of conventional statistical methods is achieved when the sample size of the data, namely “n”—the number of independent observations and subjects—is much larger than the number of candidate prediction parameters and targets, namely "p". In many genomic data analyses this situation is often completely reversed. For example, in a microarray study tens of thousands of genes' expression patterns may become the candidate prediction factors for a biological phenomenon of interest (e.g., response vs. resistance to a chemotherapeutic regimen), but the number of independent observations (e.g., different patients and/or samples) is often a few tens or hundreds at most. Due to the experimental costs and limited availability of biological materials, the number of independent samples may be even smaller, sometimes only a few. Traditional statistical methods are not designed for these circumstances and often perform very poorly; furthermore it is important to strengthen statistical power by utilizing all sources of information in large-screening genomic data.

Challenge 4: Computational limitation

We also note that no matter how powerful a computer system becomes, it is often prohibitive to solve many genomic data mining problems by exhaustive combinatorial search and comparisons [4]. In fact, many current problems in genomic data analysis have been theoretically proven to be of NP (non-polynomial)-hard complexity, implying that no computational algorithm can search for all possible candidate solutions. Thus, heuristic—most frequently statistical—algorithms that effectively search and investigate a very small portion of all possible solutions are often sought for genomic data mining problems. The success of many bioinformatics studies critically depends on the construction and use of effective and efficient heuristic algorithms, most of which are based on the careful application of probabilistic modeling and statistical inference techniques.

Challenge 5: Noisy high-throughput biological data

The next challenge derives from the fact that high-throughput biotechnical data and large biological databases are inevitably noisy because biological information and signals of interest are often observed with many other random or confounding factors. Furthermore, a one-size-fit-all experimental design for high-throughput biotechniques can introduce bias and error for many candidate targets. Therefore, many investigations in bioinformatics can be successfully performed only when such variability of genomics data are well-understood. In particular, the distributional characteristics of each data set needs to be analyzed using statistical and quality control techniques on initial data sets so that relevant statistical approaches may be applied appropriately. This preprocessing step is critical for all subsequent bioinformatics analyses, and it is sometimes difficult to reconcile dramatically different results that may stem from slightly different preprocessing procedures. While there is no easy answer for such an issue, it is important to employ consistent preprocessing procedures within each and across different analyses, with good documentation of procedures used.

Challenge 6: Integration of multiple, heterogeneous biological data for translational bioinformatics research

The last challenge is the integration of genomic data with heterogeneous biological data and associated metadata, such as gene function, biological subjects' phenotypes, and patient clinical parameters. For example, multiple heterogeneous data sets including gene expression data, biological responses, clinical findings and outcomes data may need to be combined to discover genomic biomarkers and gene networks that are relevant to disease and predictive of clinical outcomes such as cancer progression and chemo sensitivity to an anticancer compound. Some of these data sets exist in very different formats and may require combined preprocessing, mapping between data elements, or other preparatory steps prior to correlative analysis, depending on their biological characteristics and data distributions. Effective combination and utilization of the information from such heterogeneous genomic, clinical and other data resources remains a significant challenge.

III. STATISTICAL SIGNIFICANCE: FALSE DISCOVERY RATE

In this paper we review novel concepts and techniques for tackling various genomic data mining problems. In particular, because DNA microarrays and Gene Chips™ techniques have become an important tool in biological and biomedical investigations, we will focus on statistical approaches that have been applied to various microarray data analyses to overcome some of the challenges mentioned above.

In order to avoid a large number of false positive findings, the family-wise error rate (FWER) has been classically controlled for the random chance of multiple hypotheses (or candidates) by evaluating the probability that at most one false positive is included at a cutoff level of a test statistic among all candidates. However, FWER has been found to be very conservative in microarray studies, resulting in a high false-negative error rate, often very close to 100% [1]. To avoid such a pitfall, a novel concept of statistical significance, the so-called ***false discovery rate*** (FDR) and its refinement, ***q-value,*** have been suggested [2, 5] (qvalue package,)

. FDR is defined as follows. Suppose there are **M** candidates for simultaneously testing to reject the null hypothesis of no biological significance. Assume $\mathbf{M_0}$ among **M** to be the number of true negative candidates and $\mathbf{M_1}$ (=$\mathbf{M} - \mathbf{M_0}$) to be the number of true positive candidates. At a cutoff value of a test statistic or data mining tool, let R denote the number of all positives (or significantly identified candidates), V the number of false positives, and S the number of false negatives (Table 1)

Table1 Classification of the candidate hypotheses: true negative(U), false positive(V), false negative(T), true positive (S).

	Null Accept	Nullreject	Total
Null true	U	V	M_0

Alternative True	T	S	M_1
Total	W	R	M

Then, the FDR is defined as V/R if R > 0, the ratio between false positives (V) and all positive findings (R=V+S). Note that FDR is thus derived based both on the null (no significance) and alternative (significant target) distributions. In contrast, the classical p-value (or type I error), here V/M_0, and the statistical power (1 - type II error), or S/M_1, are based only on one of the null and alternative distributions. Therefore, the FDR criterion can simultaneously balance between false positives and false negatives whereas the classical p-value and power can address only one of the two errors.

The FDR evaluation has been rapidly adopted for microarray data analysis, including the widely-used SAM (Significance Analysis of Microarrays) and other approaches [1, 6]. Many different methods have been suggested for estimating FDR directly from test statistics, or indirectly from classical p-values of such statistics. The latter methods are convenient since standard p-values can be simply converted into their corresponding FDR values [5, 7] and q-value, especially the latter based on a re sampling technique. More careful FDR assessment can also be found in many other recent studies [7].

Pair wise statistical sets for genomic data

Each gene's differential expression pattern in a microarray experiment is usually assessed by (typically pairwise) contrasts of mean expression values among experimental conditions. Such comparisons have been routinely measured as fold changes whereby genes with greater than two or three fold changes are selected for further investigation. It has been found frequently that a gene that shows a high fold-change between comparison conditions might also exhibit high variability in general and hence its differential expression may not be significant. Similarly, a modest change in gene expression may be significant if its differential expression pattern is highly reproducible. A number of authors have pointed out this fundamental flaw in the fold-change based approach [1]. Thus, the emerging standard approach is based on statistical significance and hypothesis testing, with careful attention paid to reliability of variance estimates and multiple comparison issues.

The classical two-sample t-test and other traditional test statistics have been initially used for testing each gene's differential expression [6]. These classical testing procedures, however, rely on reasonable estimates of reproducibility or within-gene error, requiring a large number of replicated arrays. When a small number of replicates are available per condition, e.g., duplicate or triplicate, the use of within-gene estimates of variability does not provide a reliable hypothesis testing framework. For example, a gene may have very similar differential expression values in duplicate experiments by chance alone. Furthermore, the comparison of means can be misled by outliers with dramatically smaller or larger expression intensities than other replicates. Because of this, error estimates constructed solely within genes may result in underpowered tests for differential expression comparisons and also result in large numbers of false positives. Several approaches to improving estimates of variability and statistical tests of differential expression have thus recently emerged as follows [8–10].

IV. SIGNIFICANCE ANALYSIS OF MICROARRAY: SAM

SAM has been proposed to improve the unstable error estimation in the two-sample t-test by adding a variance stabilization factor which minimizes the variance variability across different intensity ranges [1]. Based on the observation that the signal-to-noise ratio varies with different gene expression intensities, SAM tries to stabilize gene-specific fluctuations. and is defined based on the ratio of change in gene expression to the standard deviation in the data for that gene. The relative difference d(i) in gene expression is defined as:

$$d(i)=(x_I(i) - x_U(i))/(s(i)+s0)$$

where $x_I(i)$ and $x_U(i)$ are the average expression values of gene i in states I and U, respectively. The gene-specific scatter s(i) is the standard pooled deviation of replicated expression values of the gene in the two states. To compare values of d(i) across all genes, the distribution of d(i) is assumed to be independent of the level of gene expression. However, as mentioned above, at low expression levels variability in d(i) can be high because of small values of s(i). To ensure that the variance of d(i) is independent of gene expression, a positive constant s_0 is added to the denominator. The value for s_0 is chosen to minimize the coefficient of variation, where the coefficient of variability of d(i) is computed as a function of s(i) in moving windows across all the genes.

(i) Local Pooled Error (LPE)

Based on a more careful error-pooling technique, the so-called local-pooled-error (LPE) test has also been introduced. This testing technique is particularly useful when the sample size is very small, e.g., two or three per condition. LPE variance estimates for genes are formed by pooling variance estimates for genes with similar expression intensities from replicated arrays within experimental conditions [6]. The LPE approach leverages the observations that genes with similar expression intensity values often show similar array-experimental variability within experimental conditions; and that variance of individual gene expression measurements within experimental conditions typically decreases as a (non-linear) function of intensity. LPE has been introduced specifically for analysis of small-sample microarray data, whereby error variance estimates for genes are formed by pooling variance estimates for genes with similar expression intensities from replicated arrays within experimental conditions (LPE package,)

This is possible because common background noise can often be found within each local intensity region of the microarray data. At high levels of expression intensity, this background noise is dominated by the expression intensity, while at low levels the background noise is a larger component of the observed expression intensity, which can be easily observed in the so-called AM log-intensity scatter plot of two replicated chips among three different immune conditions [6] (Figure 1)

The LPE approach controls the situation where a gene with low expression may have very low variance by chance and the resulting signal-to-noise ratio is unrealistically large. Statistical significance of the LPE-based test is evaluated as follows. First, each gene's medians m_1 and m_2 under the two compared conditions are calculated to avoid artifacts from outliers. The LPE statistic for the median (log-intensity) difference z is then calculated as:

$$z=(m_1-m_2)/s_{LPEpooled}$$

where $s_{LPEpooled}$ is the pooled standard error from the LPE-estimated baseline variances from the two conditions. The LPE approach shows a significantly better performance than two-sample t-test, SAM, and Westfall-Young's permutation tests, especially when the number of replicates is smaller than ten [6

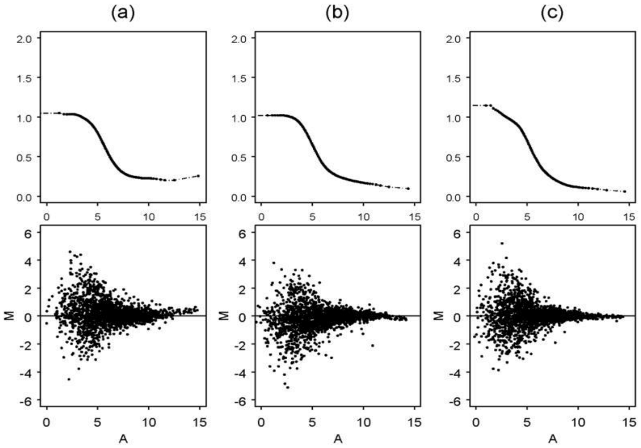

Figure.1 log intensity scatterplot of replicated chips among 3 immune conditions

V. STATISTICAL MODELING ON GENOMIC DATA:

Genomic expression profiling studies are also frequently performed for comparing complex, multiple biological conditions and pathways. Several linear modeling approaches have been introduced for analyzing microarray data with multiple conditions. For example, an ANOVA model approach was considered to capture the effects of dye, array, gene, condition, array-gene interaction, and condition-gene interaction separately on cDNA microarray data [11], and a two-stage mixed model was proposed first to model cDNA microarray data with the effects of array, condition, and condition-array interaction and then to fit the residuals with the effects of gene, gene-condition interaction, and gene-array interaction [12]. Several approaches have also been developed under the Bayesian paradigm for analyzing microarray data, including Bayesian parametric modeling [13], Bayesian regularized t-test [8], Bayesian hierarchical modeling with a multivariate normal prior [14], and Bayesian heterogeneous error model (HEM) with two error components [15]. ANOVA and HEM approaches are introduced below.

(i) ANOVA Modeling

The use of analysis of variance (ANOVA) models has been suggested to estimate relative gene expression and to account for other sources of variation in microarray data [16]. Even though the exact form of the ANOVA model depends on the particular data set, a typical ANOVA model for two-color based cDNA microarray data can be defined as

$$y_{ikg}=\mu+A_i+D_j+V_k+G_g+AD_{ij}+AG_{ig}+DG_{ig}+VG_{kg}+\varepsilon_{ijkg},$$

where y_{ijkg} is the measured intensity from array i, dye j, variety k, and gene g on an appropriate scale (typically the log scale). The generic term "variety" is often used to refer to the mRNA samples under study, such as treatment and control samples, cancer and normal cells, or time points of a biological process. The terms A, D, and AD account for the overall effects that are not gene-specific. The gene effects G_g capture the average levels of expression for genes and the array-by-gene interactions AG_{ig} capture differences due to varying sizes of spots on arrays. The dye-by-gene interactions DG_{jg} represent gene-specific dye effects. None of the above effects are of biological interest, but amount to a normalization of the data for ancillary sources of variation. The effects of primary interest are the interactions between genes and varieties, VG_g. These terms capture differences from overall averages that are attributable to the specific combination of variety k and gene g. Differences among these variety-by-gene interactions provide the estimates for the relative expression of gene g in varieties 1 and 2 by VG_{1g} – VG_{2g}. Note that AV, DV, and other higher-order interaction terms are typically assumed to be negligible and are considered together with the error terms. The error terms ε_{ijkg}'s are often assumed to be independent and normal with mean zero and a common variance. However, such a global ANOVA model is difficult to implement in practice due to its computational restriction. Instead, one often considers gene-by-gene ANOVA models like:

$$y_{ijkg}=\mu_g+A_i+D_j+V_k+AD_{ij}+VG_{kg}+\varepsilon_{ijkg}.$$

Alternatively, a two-stage ANOVA model may be used [12]. The first layer is for main effects non-specific to the gene effects:

$$y_{ijkg}=\mu+A_i+D_j+V_k+AD_{ij}+AGig+\varepsilon_{ijkg}.$$

Let r_{ijkg} be the residuals from this first ANOVA fit. Then, the second-layer ANOVA model for gene-specific effects is considered as:

$$r_{ijkg}=G_g+AG_{ig}+DG_{ig}+VG_{kg}+\nu_{ijkg}.$$

Excepting the main effects of G and V and their interaction effects, the other terms A, D, (AD), (AG), and (DG) can be considered as random effects. These within-gene ANOVA models can be implemented using most standard statistical packages, such as R (see chapter X), SAS, or SPSS.

(ii) Heterogeneous Error Model

Similarly to the statistical tests for comparing two sample conditions, the above within-gene ANOVA modeling methods are underpowered and have inaccurate error estimation in microarray data with limited replication. The heterogeneous error model (HEM) has been suggested as an alternative (HEM package)

It is based on Bayesian hierarchical modeling and LPE error-pooling-based prior constructions, with two layers of error which decompose the total error variability into the technical and biological error components in microarray data [15]. The first layer is constructed to capture the array technical variation due to many experimental error components, such as sample preparation, labeling, hybridization, and image processing:

$$y_{ijkl}=x_{ijk}+\varepsilon_{ijkl}\text{, where }\varepsilon_{ijkl}\sim\text{ iid Normal}[0,\sigma^2(x_{ijk})]\text{, where } i=1,2,\ldots,G;\ j=1,2,\ldots,C;\ k=1,2,\ldots,m_{ij};\ l=1,2,\ldots,n_{ijk}.$$

The second layer is then hierarchically constructed to capture the biological error component:

$$x_{ijk}=\mu+g_i+c_j+r_{ij}+b_{ijk}\text{, where } b_{ijk}\sim\text{ iid Normal}[0,\sigma^2{}_b(ij)].$$

Here, the genetic parameters are for the grand mean (shift or scaling) constant, gene, cell, interaction effects, and the biological error; the last error term varies and is heterogeneous for each combination of different genes and conditions. Note that the biological variability is individually assessed for discovery of biologically-relevant expression patterns. The HEM approach shows a significantly better performance than standard ANOVA methods, especially when the number of replicates is small (Figure 2)

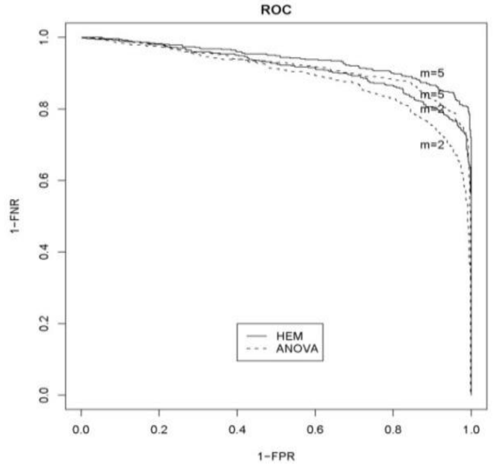

Figure 2: ROC curves from HEM (solid lines) and ANOVA (dotted lines) models with two and five replicated arrays; The horizontal axis is 1 - FPR (false positive error rate) and the vertical axis is 1 – FNR (false negative error rate).

VI. CONCLUSION

We believe that there are several unprecedented challenges in the analysis of why these genomic data . Data mining approaches are a promising direction for future work. Clustering , classification models and genomic pathway models can be the other modeling techniques. Genomic expression signatures are highly relevant to the results of statistical modeling and the significant analysis of microarray. Even though many individual genes' expression values in such networks are often variable and noisy, the whole gene networks have been found to be quite consistent in their overall expression patterns . Genome-wide RNA expression profiling techniques such as microarrays and GeneChips™ have been dramatically improved in recent years, so that the expression patterns of the entire human genome can now be accurately and cost-effectively measured on patient samples. In fact, microarray RNA profiling is one of the most accurately quantifiable and comprehensive profiling biotechnologies among all current high-throughput biotechniques,

REFERENCES

[1] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A.* 2001;98(9):5116–21. [PubMed]

[2] Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A.* 2003;100(16):9440–5. [PubMed]

[3] Hastie T, et al. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 2000;1(2):RESEARCH0003. [PubMed]

[4] Soukup M, Cho H, Lee JK. Robust classification modeling on microarray data using misclassification penalized posterior. *Bioinformatics.* 2005;21(Suppl 1):i423–i430. [PubMed]

[5] Benjamini Y, et al. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res.* 2001;125(1–2):279–84. [PubMed]

[6] Jain N, et al. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics.* 2003;19(15):1945–51. [PubMed]

[7] Jain N, et al. Rank-invariant resampling based estimation of false discovery rate for analysis of small sample microarray data. *BMC Bioinformatics.* 2005;6:187. [PubMed]

[8] Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics.* 2001;17(6):509–19. [PubMed]

[9] Efron B, Tibshirani R. Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol.* 2002;23(1):70–86. [PubMed]

[10] Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *J Comput Biol.* 2000;7(6):819–37. [PubMed]

[11] Kerr MK, Churchill GA. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc Natl Acad Sci U S A.* 2001;98(16):8961–5. [PubMed]

[12] Wolfinger RD, et al. Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol.* 2001;8(6):625–37. [PubMed]

[13] Newton MA, et al. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol.* 2001;8(1):37–52. [PubMed]

[14] Ibrahim JGaCM-H, Gray RJ. Bayesian Models for Gene Expression with DNA Microarray Data. *Journal of American Statistical Association.* 2002;97:88–99.

[15] Cho H, Lee JK. Bayesian hierarchical error model for analysis of gene expression data. *Bioinformatics.* 2004;20(13):2016–25. [PubMed]

[16] Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarray data. *Genet Res.* 2001;77(2):123–8. [PubMed]

[17] Lee JK, et al. Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells. *Genome Biol.* 2003;4(12):R82. [PubMed]

ABOUT AUTHORS

[1] Mrs. S.Jessica Saritha is an Assistant Professor in the Department of CSE , JNTUA College of Engineering Pulivendula city in Andhra Pradesh. She has done B.Tech from JNTU Anantapur , , M.Tech from JNTU Kakinada and perusing PhD from JNTU Hyderabad in Telangana State. Her research reas are Data mining and bioinformatics

[2] Professor P.Govindarajulu is a retired professor from Sri Venkateswara University Tirupathi , Andhra Pradesh He served at several portfolios at the University while in service and currently the honorary director . He has obtained hid M.Tech from IIT Chennai. and PhD from IIT Mumbai. His research interests are data mining and databases.