Semi Supervised Based Framework for Gene Data Analysis Using SVM Classification and Random Forest Approach

VINOTHA.S[1], SELVAM AMALRAJ.M[2]
[1]PG Student, Dept of CSE, Bharathiyar College of Engineering and Technology, Karaikal.
[2]Assistant Professor, Dept. of CSE, Bharathiyar College of Engineering and Technology, Karaikal.

Abstract— Microarray technology is one of the important biotechnological means that allows recording the expression levels of thousands of genes simultaneously within a number of different samples. A microarray gene expression data set can be represented by an expression table, where each row corresponds to one particular gene, each column to a sample, and each entry of the matrix is the measured expression level of a particular gene in a sample, respectively. An important application of microarray gene expression data in functional genomics is to classify samples according to their gene expression profiles. Among the large amount of genes presented in gene expression data, only a small fraction of them is effective for performing a certain diagnostic test. However, for most gene expression data, the number of training samples is still very small compared to the large number of genes involved in the experiments. When the number of genes is significantly greater than the number of samples, it is possible to find biologically relevant correlations of gene behavior with the sample categories or response variables. Hence, one of the major tasks with the gene expression data is to find groups of co-regulated genes whose collective expression is strongly associated with the sample categories or response variables. So implement feature subset selection approach to reduce dimensionality, removing irrelevant data and increase diagnosis accuracy and presents learning method which is able to group genes based on their interdependence so as to mine meaningful patterns from the gene expression data using Spatial EM algorithm. It can be used to calculate spatial mean and rank based scatter matrix to extract relevant patterns and further implement KNN (K- nearest neighbor classification) approach to diagnosis the diseases. An important finding is that the proposed semi supervised clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability. An optimum random forest based algorithm is proposed for the analysis.

Keywords- Data warehouse, Random Forest Algorithm, Knowledge Discovery in Data (KDD), Data Mining, Micro Array analysis.

1. INTRODCUTION

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

Operational or transactional data such as, sales, cost, inventory, payroll, and accounting Nonoperational data, such as industry sales, forecast data, and macro economic data Meta data - data about the data itself, such as logical database design or data dictionary definitions . In computing, a data warehouse (DW or DWH) is a database used for reporting and data analysis. It is a central repository of data which is created

by integrating data from multiple disparate sources. Data warehouses store current as well as historical data and are commonly used for creating trending reports for senior management reporting such as annual and quarterly comparisons. The data stored in the warehouse are uploaded from the operational systems (such as marketing, sales etc., shown in the figure to the right). The data may pass through an operational data store for additional operations before they are used in the DW for reporting. The typical ETL-based data warehouse uses staging, integration, and access layers to house its key functions. The staging layer or staging database stores raw data extracted from each of the disparate source data systems. The integration layer integrates the disparate data sets by transforming the data from the staging layer often storing this transformed data in an operational data store (ODS) database. The integrated data are then moved to yet another database, often called the data warehouse database, where the data is arranged into hierarchical groups often called dimensions and into facts and aggregate facts. The combination of facts and dimensions is sometimes called a star schema. The access layer helps users retrieve data.

A data warehouse constructed from integrated data source systems does not require ETL, staging databases, or operational data store databases. The integrated data source systems may be considered to be a part of a distributed operational data store layer. Data federation methods or data virtualization methods may be used to access the distributed integrated source data systems to consolidate and aggregate data directly into the data warehouse database tables. Unlike the ETL-based data warehouse, the integrated source data systems and the data warehouse are all integrated since there is no transformation of dimensional or reference data. This integrated data warehouse architecture supports the drill down from the aggregate data of the data warehouse to the transactional data of the integrated source data systems.

1.1 Different levels of analysis

Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset where k 1). Sometimes called the k-nearest neighbor technique.

Rule induction: The extraction of useful if-then rules from data based on statistical significance.

Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

Classification

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, make the software that can learn how to classify the data items into groups. For example, can apply classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, divide the employee's records into two groups that are "leave" and "stay". And then can ask data mining software to classify the employees into each group.

2. RELATED WORK

High breakdown mixture discriminant analysis

The classification rules depend on the unknown parameters, which are to be estimated from the training data. In the presence of a number of outlying observations in the training data, the estimates of the unknown parameters can be

unstable due to the undue influence of these atypical observations. High breakdown estimation is a procedure designed to remove this cause of concern, by producing estimators that are robust to serious distortion by outliers, eliminating the influence of such atypical observations. However, it is an important fact that in discriminate analysis, not only are the outliers a concern but also inliers. In the K-means clustering, the outliers for one group might be the inliers for others affecting the classification performance, while in case of mixtures of distributions, this situation may be even worse. The conventional maximum likelihood estimators are affected by the presence of outliers, and so break down. These non-robust estimators influence the discriminate function, leading to the poor classification. The mda approach resulted in the smallest errors of misclassification. It is because the mda approach with maximum likelihood estimators works well within the set of assumptions on which it is based.

Depth-Based Novelty Detection and its Application to Taxonomic Research

The job of discovering and describing new species falls on taxonomists. The science of taxonomy has also been suffering from dwindling numbers of experts over the past few decades. Moreover, the pace of taxonomic research, as traditionally practiced, is very slow. In recognizing a species as new to science, taxonomists use a gestalt recognition system that integrates multiple characters of body shape, external body characteristics, and pigmentation patterns. They then make careful counts and measurements on large numbers of specimens from multiple populations across the geographic ranges of both the new and closely related species, and identify a set of external body characters that uniquely diagnoses the new species as distinct from all of its known relatives. The process is laborious and can take years or even decades to complete, depending on the geographic range of the species and believe that the pace of data gathering and analysis in taxonomy can be greatly increased through the integration of machine learning and data mining techniques into taxonomic research and tackle one of the most important and challenging research objectives in taxonomy new species discovery and develop a novelty detection framework that avoids the above limitation of spatial depth. Specifically, introduce a new depth function, kernelized spatial depth (KSD), which defines the spatial depth in a feature space induced by a positive definite kernel.

Outlier Detection with the Kernelized SpatialDepth Function

Analyze a novel outlier detection framework based on the notion of statistical depths. Outlier detection methods that are based on statistical depths have been studied in statistics and computational geometry. These methods provide a center-outward ordering of observations. Outliers are expected to appear more likely in outer layers with small depth values than in inner layers with large depth values. Depth-based methods are completely data-driven and avoid strong distributional assumption. Moreover, they provide intuitive visualization of the data set via depth contours for a low dimensional input space. However, most of the current depth-based methods do not scale up with the dimensionality of the input space. For example, finding peeling and depth contours, in practice, require the computation of d dimensional convex hulls. Because each observation from a data set contributes equally to the value of depth function, spatial depth takes a global view of the data set. Consequently the outliers can be called as "global" outliers. Nevertheless, many data sets from real-world applications exhibit more delicate structures that entail identification of outliers relative to their neighborhood, i.e., "local" outliers and develop an outlier detection framework that avoids the above limitation of spatial depth.

Nonparametric Depth-Based Multivariate Outlier Identifiers and Masking Robustness Properties

An outlier identifier must; of course, be itself robust in the presence of the outliers it is supposed to identify. As a key relevant robustness criterion, introduce the masking breakdown point (MBP), which measures the fraction of sample allowed to be contaminants without some extreme outlier becoming "masked", i.e., misidentified as a non outlier and use replacement contamination. The approach adapts a notion introduced and using Mahalanob is distance out lyingness with the contaminated normal model and addition type contamination. (While not identical, replacement and addition breakdown points are equivalent as measures of robustness performance, although differing in intuitive appeal. In particular, derive and compare MBPs for four affine invariant outlyingness functions,based on the well-established Mahalanob is distance, half space, and projection depths, and on a new "Mahalanob is spatial" depth recently treated in Serfling. The latter has atransformation-

retransformation representation in terms of the well-known "spatial" outlyingness, which is only orthogonally invariant.

3. PROPOSED SYSTEM

In proposed, implement Spatial EM algorithm for analyzing microarray datasets. It is used to identify cluster location from group gene datasets by utilizing robust location and scatter estimators in each M-step. Able to represent arbitrarily complex structure of data. Another common technique for robust fitting of mixtures is to update the component estimates on the M-step of the EM algorithm by some robust location and scatter estimates. Mestimator has been considered. It used minimum covariance determinant (MCD) estimator for cluster analysis. Recommended the use of S estimator. In this paper, we propose to apply spatial rank based location and scatter estimators. They are highly robust and are computationally and statistic ally more efficient than the above robust estimators .We develop a Spatial-EM algorithm for robust finite mixture learning. Based on the Spatial-EM, supervised outlier detection and unsupervised clustering methods are illustrated and compared with other existing techniques.

EM Algorithm

EM estimation has been proved to converge to maximum likelihood estimation (MLE) of the mixture parameters under mild conditions. The above simple implementation makes Gaussian mixture models popular. However, a major limitation of Gaussian mixture models is their lack of robustness to outliers. This is easily understood because maximization of likelihood function under an assumed Gaussian distribution is equivalent to finding the least-squares solution, whose lack of robustness is well known. Moreover, from the perspective of robust statistics, using sample mean and sample covariance of each component in the M-step causes the sensitivity problem because they have the lowest possible breakdown point. Here the breakdown point is a prevailing quantitative robustness measure proposed by Donohue and Huber. Roughly speaking, the breakdown point is the minimum fraction of "bad" data points that can render the estimator beyond any boundary. It is clear to see that one point kxk!1 is enough to ruin the sample mean and sample covariance matrix. Thus, their breakdown point is 1=n. As a robust alternative, mixtures of t-distributions have been used for modeling data that have wider tails than Gaussian's observations.

MODULES:

Datasets Acquisition

In this module, upload the datasets. The dataset may be microarray dataset. A microarray database is a repository containing microarray gene expression data. The key uses of a microarray database are to store the measurement data, manage a searchable index, and make the data available to other applications for analysis and interpretation. Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time.

Median Estimation

To tackle the effect of outliers in cluster analysis to consider the Spatial EM clustering which replaces the squared Euclidean distances in the objective function of the k-means clustering with the absolute Euclidean distances. In spatial EM, can analyze coverage of the data before clustering begins. And propose an algorithm, which modifies the nearest centroid sorting and the transfer algorithm, of the spatial medians clustering. It has two distinct phases: one of transferring an object from one cluster to another and the other of amalgamating the single member cluster with it's the nearest cluster. Given a starting partition, each possible transfer is tested in turn to see if it would improve the value of clustering criterion. When no further transfers can improve the criterion value, each possible amalgamation of the single member cluster and other clusters is tested. The amalgamation of the single member cluster should be executed with the detachment of an object which is far from its cluster centroid when it is found to be beneficial. When no further amalgamations give an improvement, the transfer phase is reentered and continued until no more transfers or amalgamations can improve the clustering criterion value. In this module,

can calculate the mean values for each gene features.

Rank based scatter

In this module, can create scatter matrix based on median values that are derived by clustering algorithm. Then construct scatter matrix and reflecting as the within-cluster scatter, the between-cluster scatter and their summation — the total scatter matrix. The determinant of a scatter matrix roughly measures the square of the scattering volume. And minimizing this measure is equivalent to both minimizing the intra-cluster scatter and maximizing the inter-cluster scatter. Based on scatter matrix, classification is performed in following modules. Mixture model-based clustering is one of the most popular and successful unsupervised learning approaches. It provides a probabilistic (soft) clustering of the data in terms of the fitted posterior probabilities (Tji's) of membership of the mixture components with respect to the clusters. An outright (hard) clustering can be subsequently obtained by assigning each observation to the component to which it has the highest fitted posterior probability of belonging.

Random Forest Approach

In decision tree algorithm of Random Forest, the tree is constructed dynamically with online fitting procedure. A random forest is a substantial modification of bagging .Each tree of Random Forest is grown can be explained as follows: Suppose training data size containing N number of records, then N records are sampled at random but with replacement, from the original data, this is known as bootstrap sample along with M number of attributes. This sample will be used for the training set for growing the tree. If there are N input variables, a number n N is selected such that at each node, n variables are selected at random out of N and the best split on these m attributes is used to split the node. The value of m is held constant during forest growing. The decision tree is grown to the largest extent possible. A tree forms "inbag" dataset by sampling with replacement member from the training set.

Evaluation criteria

In this module, the performance of the proposed semi-supervised algorithm is extensively compared with that of some existing supervised and unsupervised gene clustering and gene selection algorithms. To analyze the performance of different algorithms, the experimentation is done on

microarray gene expression data sets. The major metrics for evaluating the performance of different algorithms are the class separability index and classification accuracy of K-nearest neighbor rule. The proposed system provide improved accuracy rate in gene classification.

4. CONCLUSION AND FUTURE WORK

Recent DNA microarray technologies have made it possible to monitor transcription levels of tens of thousands of genes in parallel. Gene expression data generated by microarray experiments offer tremendous potential for advances in molecular biology and functional genomics. This paper reviewed both classical and recently developed clustering algorithms, which have been applied to gene expression data, with promising results. The proposed semi-supervised spatial EM clustering algorithm is based on measuring mean values and scatter matix using the new quantitative measure, whereby redundancy among the attributes is removed. The clusters are then refined incrementally based on sample categories. The performance of the proposed algorithm is compared with that of existing supervised EM gene selection algorithm with accuracy rate. An important finding is that the proposed semi-supervised clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability.

FUTURE WORK:

In future, can extend the work to implement this concept with multi classification. The multi classification is used to identify the diseases with various severity levels and recommend the prescription details.

REFERENCES

1] S. Bashir and E. M. Carter, "High breakdown mixture discriminant analysis," J. Multivariate Anal., vol. 93, no. 1, pp. 102–111, 2005.

[2] C. Biernacki, G. Celeux, and G. Govaert, "An improvement of the NEC criterion for assessing the number of clusters in a mixture model," Pattern Recognit. Lett., vol. 20, pp. 267–272, 1999.

[3] B. Brown, "Statistical uses of the spatial median," J. Roy. Stat. Soc., B, vol. 45, pp. 25–30, 1983.

[4] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler, "Knowledge-based analysis of microarray gene expression fidata by using support vector machines,"

Proc. Nat. Acad. Sci., vol. 97, no. 1, pp. 262–267, 2000.

[5] N. A. Campbell, "Mixture models and atypical values," Math. Geol., vol. 16, pp. 465–477, 1984.

[6] G. Celeux and G. Soromenho, "An entropy criterion for assessing the number of clusters in a mixture model," Classification J., vol. 13, pp. 195–212, 1996.

[7] Y. Chen, Bart H. Jr, X. Dang, and H. Peng, "Depth-based novelty detection and its application to taxonomic research," in Proc. 7th IEEE Int. Conf. Data Mining, Omaha, Nebraska, 2007, pp. 113–122.

[8] Y. Chen, X. Dang, H. Peng, and H. Bart Jr., "Outlier detection with the kernelized spatial depth function," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 2, pp. 288–305, Feb. 2009.

[9] Y. Chueng, "Maximum weighted likelihood via rival penalized EM for density mixture clustering with automatic model selection," IEEE Trans. Knowl. Data Eng., vol. 17, no. 6, pp. 750–761, Jun. 2005.

[10] X. Dang and R. Serfling, "Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties," J. Stat. Inference Planning, vol. 140, pp. 198–213, 2010.