International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume III, Issue XII, December- 2016

Criticality Scores and Optimized Support Vector Machine for Cancer Risk Assessment

Ms. C. Nithya M.C.A, MPhil (Research Scholar)
Ms. S.N.SANTHALAKSHMI M.C.A.,M.Phil.,M.Ed, *Assistant Professor,*
Department of Computer Science,
Nandha Arts and Science College, Erode, Tamilnadu, India

Abstract

The gene expression data elements are managed under the micro array data structures. Genetic behaviors of the cancer patients are reflected in the gene expressions. Cancer risk levels can be discovered using the gene data values. Cancer patients are assigned with four risk levels such s low, medium, high and very high risk levels. Treatment and medicine decisions are made with reference to the risk level of the patients. The risk assessment operations are carried out on the Leukemia patient diagnosis data values.

Machine learning methods are applied for the cancer risk assessment process. High dimensional data classification methods are adapted for the gene expression risk assessment process. The gene selection and classification tasks are carried out using the Hybrid Algorithm. The gene selection is performed with Genetic Algorithm (GA) and Simulated Annealing (SA) algorithms. The classification process is initiated on the selected gene data values. The Feature Selection based Support Vector Machine (FS-SVM) algorithm is applied for the classification process.

The cancer risk assessment process is performed using the criticality scores and Optimized Support Vector Machine methods. The Optimized Feature Selection based Support Vector Machine (OFS-SVM) algorithm analyzes the cancer gene expression values for risk discovery process. Multi objective genetic algorithm with mixed mutation model is adapted for the gene selection process. The gene feature selection process is improved with criticality scores. The classification process is performed on the selected gene feature values. Cancer risk levels are derived from the classification results.

1. Introduction

Nowadays there is huge amount of data being collected and stored in databases everywhere across the globe. The tendency is to keep increasing year after year. It is not hard to find databases with Terabytes of data in enterprises and research facilities. That is over 1,099,511,627,776 bytes of data. There is invaluable information and knowledge "hidden" in such databases; and without automatic methods for extracting this information it is practically impossible to mine for them. Throughout the years many algorithms were created to extract what is called nuggets of knowledge from large sets of data. There are several different methodologies to approach this problem: classification, association rule, clustering, etc.

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute [6]. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is. For example, in a medical database the training set would have relevant patient information recorded previously, where the prediction attribute is whether or not the patient had a heart problem.

2. Related Work

There has been a lot of prior works on spectral clustering for gene expression data. For example, Braun et al. [1] explored the Partition Decoupling Method (PDM)-based on iterated spectral clustering, and applied it to three cancer gene expression data sets. Qiu and Plevritis [8] proposed a new approach, named as spectral analysis for class discovery, and classification, to identify biologically meaningful subclasses from cancer data sets. Kluger et al. investigated how to apply the spectral biclustering approach to identify the underlying structure of cancer data sets. De Souto et al [13] performed a comparative study on different clustering approaches including spectral clustering for class discovery from cancer gene expression profiles. Higham et al. compared the effect of normalized and unnormalized spectral clustering on microarray data sets. However, few of these works consider how to incorporate the spectral clustering approach into the consensus clustering framework to improve the accuracy, robustness, and stability of the final results.

Consensus clustering approaches, also referred to as cluster ensemble approaches, are gaining more and more attention, due to its useful applications in the areas of bioinformatics, pattern recognition, data mining, machine learning, and so on. Consensus clustering approaches have the powerful capability to perform fusion of multiple partitions from different data sources and improve the robustness and stability of traditional single clustering algorithms. Consensus clustering approaches mainly consist of two stages: the clustering ensemble generation stage and the consensus aggregation stage. The main focus of the first stage is to generate a set of clustering solutions which are as diverse as possible, while the second stage focuses on finding a good consensus aggregation of the individual clustering solutions which can improve the accuracy of the final result.

In order to successfully discover different cancer types from the gene expression profiles, researchers apply different consensus clustering approaches to perform cluster analysis on these data sets. There are a number of different consensus clustering approaches which can be divided into two categories based on their emphases. The approaches in the first category designed new consensus clustering techniques, and applied them to cancer gene expression profiles. For example, Duboit et al. applied a prediction-based resampling consensus clustering approach named Clest to perform class discovery from four cancer microarray data sets. Smolkin and Ghosh proposed a new consensus clustering approach based on the random subspace technique and a cluster stability score to identify the number of clusters from cancer data sets.

Monti et al. adopted two kinds of consensus clustering approaches, which are the Self-Organizing Map (SOM)-based consensus clustering approach and the hierarchical clustering-based consensus clustering approach, to discover the underlying structure from cancer gene expression data. Yu and Wong [2] proposed a Graph-Based Consensus Clustering (GCC) algorithm to identify cancer subtypes form gene expression profiles. They [3] also designed a new cluster ensemble approach based on the perturbation technique and the neural gas algorithm to perform clustering analysis on cancer data sets, and proposed a new cluster validity index to identify the number of cancer subtypes. Smyth and Coomans [4] presented a weighted consensus clustering approach by considering the relative accuracy of each clustering solution, and applied it to perform cluster analysis on cancer gene expression profiles. Valentini and Bertoni introduced the randomized map-based consensus clustering approach, and applied it to discover the meaningful clusters from the gene expression profiles. They also proposed a consensus clustering approach based on the Model Order Selection by Randomized Maps (MOSRAM) to discover the subclasses of patients from cancer data. Simpson et al. [9] proposed the merged consensus clustering approach based on the resampling statistics technique to perform class discovery on cancer data sets. Iam-on et al. [10] explored how to apply the consensus clustering approach to perform robust multiscale clustering analysis on microarray data. Grotkjaer et al. designed a link-based cluster ensemble method to perform cluster analysis on cancer data sets, and achieved good results.

The approaches [12] in the second category focus on introducing new theories to the consensus clustering framework. For example, Avogadri and Valentini [11] adopted fuzzy theory and proposed a random projection- based fuzzy clustering ensemble approach to discover the biological meaningful clusters from cancer gene expression profiles. Yu and Wong [7] introduced the knowledge learning framework, and designed a corresponding knowledge- based cluster ensemble approach which incorporates prior knowledge from experts into the cluster ensemble framework to assign samples to their corresponding cancer types. Bertoni and Valentini incorporated the Bernstein's inequality into the consensus clustering framework and applied it to identify cancer subtypes from gene expression data, such as leukemia, lymphoma, adenocarcinoma, and melanoma.

There are also a number of excellent reviews on existing consensus clustering approaches. For example, Handl et al. performed a review of data perturbation-based consensus clustering approaches and pointed out that these techniques are useful for characterizing and evaluating the stability of clustering results. Our proposed approaches belong to the first category. Compared with the previous consensus clustering approaches, the main features of SC3 and SC2Ncut include 1) the adoption of the spectral clustering algorithm to select representative genes in the gene dimension, which is able to reduce the effect of noisy genes; 2) the adoption of hybrid distance functions to increase the diversity of the cluster ensemble; and 3) the adoption of the normalized cut algorithm to partition the consensus matrix and improve the accuracy of the final results.

Class discovery consists of two stages: 1) identifying the number of clusters in a data set with a suitable cluster validity index, and 2) assigning data samples to their correct clusters with a suitable clustering algorithm. There are a number of cluster validity indices that could be used in the first stage, such as the Akaike Information Criterion (AIC), the Minimum Description Length criterion (MDL), the Bayesian Information Criterion (BIC), the Silhouette Index (SI), the Davies-Bouldin index (DBI), the Dunn index (DI) the Gap statistic and so on. Different kinds of clustering algorithms could be applied to microarray data at the second stage of class discovery, such as the self-organizing feature maps, the two-way clustering method, hierarchical clustering, spectral clustering, consensus clustering approaches and so on. Our proposed triple spectral clustering-based consensus clustering framework mainly focuses on the improvement of the second stage of class discovery, which assigns data samples to their corresponding clusters as correctly as possible.

3. Feature Selection based Support Vector Machine (FS-SVM) Technique

Deoxyribo Nucleic Acid (DNA) acts as a template for making copies of itself and also as a blueprint for a molecule called Ribo Nucleic Acid (RNA). The process of transcribing a gene's DNA sequence into RNA is called gene expression. A gene's expression level indicates the approximate number of copies of that gene's RNA produced in a cell and it is correlated with amount of the corresponding proteins made. Microarray is the technology for measuring the expression levels of tens of thousands of genes in parallel in a single chip. Each chip is about 2cm by 2cm and microarrays contain up to 6000 spots.

Different Microarray technologies include Serial Analysis of Gene Expression (SAGE), nylon membrane and illumina bead array. Thus, microarrays offer an efficient method of gathering data that can be used to determine the expression pattern of thousands of genes. High dimensionality of gene expression data is a big challenge in most classification problems. Large number of features against small sample size and redundancy in expressed data are the main two reasons which lead to poor classification accuracy. Subsequently dimension reduction is essential to classification. Support Vector Machine (SVM) is a supervised computer learning technique used for data classification. SVM's have been performing well in evaluating microarray expression data. It performs classification on data by placing an optimal hyper plane which maximizes the functional margin.

The system uses hybrid feature selection technique which is a combination of SVM-RFE and Based Bayes Filter (BBF) for gene selection and sequential minimal optimization algorithm for training SVM classification method. The experiments have been conducted on publicly available leukemia data set. The results bring new insights on feature selection and achieve better accuracy than classification methods.

The gene expression data values are used to analyze the diseases. The cancer diseases are analyzed using the gene expression data values. The hybrid algorithm is applied to perform the cancer classification process. The gene selection and classification tasks are carried out under the hybrid algorithm. Feature selection or gene selection operations are carried out to identify the gene attributes for the classification process. Filtering methods are applied for the gene selection process. The Genetic Algorithm (GA) is integrated with the Simulated Annealing (SA) method for the gene selection process. The Support Vector Machine technique is used for the classification process. The Feature Selection based Support Vector Machine (FS-SVM) is applied for the gene selection and classification process. The cancer classification task classifies the gene expression data into two classes' benign and malignant categories. The following drawbacks are identified from the existing system. Subset selection is not optimized, Computational complexity is high, Prediction accuracy is low and Process time is high.

4. Optimized Feature Selection based Support Vector Machine (OFS-SVM) Technique

The classification techniques are used to predict disease information from the diagnosis data values. Gene expression is micro array data values to maintain the genome details. The gene expression data is a high dimensional data value. High dimensional data process is a complex task. Data cleaning and feature selection methods are applied to improve the process results. Gene expressions are used to predict the cancer disease levels. Cancer prediction and severity classification tasks are carried out on the gene expression data values. The classification process is divided into two tasks. They are training process and testing process. The training phase learns the class patterns from the labeled transactions. The testing phases identify the class values using the learned patterns.

The hybrid algorithm is used to perform the cancer classification process. Gene selection and cancer classification operations are gene expression data values. The Feature Selection based Support Vector Machine (FS-SVM) scheme is used to handle the gene selection and classification operations. The feature selection process is applied to select the suitable gene attribute values. The filtering methods are applied for the gene selection process. The Genetic Algorithm (GA) and Simulated Annealing (SA) methods are combined to perform the gene selection process. The Multi Objective Genetic Algorithm is used to improve the gene selection process. The mutation models are used to rotate the data levels in the Genetic Algorithm. The mixed mutation model integrates the uniform mutation, boundary mutation and transformed mutation techniques. The gene selection is enhanced with mixed mutation models.

The criticality score is used for the feature selection process. The gene attributes are analyzed to estimate the criticality score values. The feature selection is performed with criticality core values. The optimized feature selection is carried out with the support of the criticality score values. The classification process is performed with the Support Vector Machine method. The Optimized Feature Selection based Support Vector Machine (OFS-SVM) method is adapted to perform the gene selection and cancer classification operations. The multi class classification process is carried out on the cancer gene expression data values. Cancer severity levels are estimated on the Leukemia cancer gene expression data values.

5. Cancer Risk Assessment Framework

The gene expression data are maintained under the micro array data structure. Cancer gene values are maintained under the gene expression. The gene selection or feature selection methods are used to filter the gene values. The classification methods are applied to classify the cancer patients. The filtering techniques are adapted to perform the gene selection process. The Genetic Algorithm (GA) and Simulated Annealing (SA) methods are combined for the gene selection process. The hybrid algorithm is used for the gene selection and classification process. Feature Selection based Support Vector Machine (FS-SVM) technique is used for the feature selection and classification process. The classification and feature selection process is improved with Optimized Feature Selection based Support Vector Machine (OFS-SVM) scheme.

5.1. Genetic Algorithm

In this method, a search is conducted in the space of genes, evaluating the goodness of each gene subset by the estimation of the accuracy percentage of the specific classifier to be used, training the classifier only with the found genes. It is claimed that this approach obtains better predictive accuracy estimates than the previous approach. A common drawback in this method is that they have a higher risk of over fitting than filter techniques and are very computationally intensive. In contrast, it incorporate the interaction between genes selection and classification model, which make them unique.

5.2. Hybrid Gene Selection Technique

This new ensemble approach is the combination of SVM-RFE and BBF. SVM-RFE yields good performance on classification but lacks on poor seperability in redundant class labels. BBF avoids redundant class labels in selection. Both combined achieves comparable performance. The schematic view of overall process carried. The feature selection from dataset is performed with SVM-RFE and BBF. After selection it is undergone with classifier for training. Finally evaluation carried with testing data.

5.2.1. Feature Selection Method

The recursive elimination procedure of SVM-RFE [5] is implemented as follows:

1. Start: ranked set *R*=[ ]; picked feature subset *S*=[1,…, *d*].

2. Repeat until all features in subset gets ranked: a. Train the features with SVM from set S as input variables.
b. Calculate the weight vector for each feature.
c. Calculate the ranking score for features in set *S*.
d. Identify the feature with the smallest ranking score.
e. Update.
f. Eliminate smallest ranking feature.

3. Result: Ranked feature set *R*[ ].

After ranking genes by SVM-RFE the system eliminates redundancy by applying BBF. First the relevant candidate genes are selected by a criterion function and second the criterion controlling the upper bound of the Bayes error is applied to the relevant candidate genes in order to remove the redundant genes. This method can effectively perform gene selection with reasonably low classification error rates and a small number of selected genes. This not only obtains a small subset of informative genes for classification analysis, but also provides a balance between selected gene set size and classification accuracy.

5.2.2. Classification with SVM

SVM is a data mining technique which classifies data in an intelligent manner. The SVM learns itself by separating data with a plane on a given training data and regression rules from data. SVM was first outlined by Vapnik *et al.* from statistical learning methods in the 1960 for classifying the data. SVM classifies data in large data sets by identifying a linear or non-linear separating surface in the input space of a data set. The separating surface depends only on the subset of the original data known as a set of support vectors. A SVM classifies data by placing one or more planes on data such that it achieves good classification results. A good result on separation is achieved by plane that has the largest distance to the nearest data points of any class, called functional margin. If this functional margin is large, then the generalization error of the classifier will be small and vice versa.

5.3. Criticality Scores

Consider a training data set T_r with m data instances, each instance having n attributes denoted as A_j (jε{1, 2, . . . , n}). The underlying assumption is that all attributes are numeric and not categorical. From T_r, form a neighborhood N, by choosing a data instance D_i as a center and finding a group of points that belong to the same class as D_i and lying within a distance R from D_i. For simplicity, let us say that the neighborhood N is comprised of d data instances. The selection of parameters R and D_i used in forming a neighborhood N. First, a classification model M_0 is generated by applying a classification algorithm C to the training data set T_r. Using the classification model M_0, one can predict the class labels for the different data instances in question. For the d instances in neighborhood N, consider an attribute A_j. Also, for the d instances, the attribute A_j can be increased or decreased in magnitude.

A parameter denoted by δ_j is used for this and δ_j varies for different attributes in neighborhood N. After increasing A_j by an extent δ_j for just the d instances, the classification model M_0 for the new class labels for the d instances is queried. The average number of data instances that have switched classes in neighborhood N is computed and is denoted as W^{-}_{-}. If all the data instances in N switch classes, then one can infer that N is very sensitive to changes with respect to attribute A_j. The same test is applied on N by decreasing A_j by the same extent δ_j and find W_j^{+} by querying the classification model M_0 for the new class labels. For the attribute A_j, the average of W^{-}_{-} and W^{-}_{j} is computed to get w_j. Repeating this process for all n attributes, the average of the w_j scores is computed as the CR_{score} for the neighborhood N. Formally, the critical score is defined as

$$CR_{score} = \frac{\sum_{j=1}^{n} w_j}{n} \qquad - \qquad (1)$$

where: $w_j = \frac{w_j^{+} , w_j^{-}}{2}$, $W_j^{+} = \frac{d_j^{+}}{d}$ and $W_j^{-} = \frac{d_j^{+}}{d}$.

Using the description on how the CR_{score} is calculated, the algorithm GetNuggetScore is developed. The computational complexity of the algorithm is derived as follows: Deriving the model M_0 is dependent on the complexity of the chosen classification algorithm (C). The complexity of the classification algorithm is denoted as t(C). Each attribute A_j is analyzed by checking if increasing or decreasing the values of the attributes by an extent δ_j, switches the class label. Hence, for each attribute, the model M_0 is queried twice. There are d data instances in N and for each attribute there are 2 × d queries. Since there are n attributes, the complexity of the for-loop is O(dn). When $d \ll n$, the complexity of the for-loop becomes $\approx$O(n). The total complexity of the algorithm is O(t(C) + dn) ($\approx$ O(t(C) + n) when $d \ll n$).

5.4. Cancer Risk Assessment using OFS-SVM

The cancer classification process is build to classify the cancer patients using the gene expression data values. The feature selection and classification operations are carried out on the gene expression data values. The gene expression data is a high dimensional data value. The gene selection is also referred as feature selection process. The filtering methods are applied for the gene selection process. The Genetic Algorithm (GA) and Simulated Annealing (SA) methods are combined for the gene selection process. The mixed mutation mechanism is applied in the Genetic Algorithm operations. The multi objective genetic algorithm is used in the feature selection process. The optimal features are selected with the improved genetic algorithm mechanism.

The feature selection process is enhanced with criticality score based filtering method. The criticality score is estimated for all gene attributes. The gene attributes are selected with reference to the criticality score values. The Optimized Feature Selection based Support Vector Machine (OFS-SVM) scheme uses the optimal features for the classification process. The multi class categorization process is carried out in the system. Gene selection and ranking operations are carried out before the classification process. The class intervals are identified using the criticality score. Cancer severity levels are estimated in the classification process.

7. Conclusion and Future Enhancement

The cancer gene expressions are high dimensional data values. Micro arrays are used to maintain the gene expression data values. Classification methods are applied to categorize the gene expression data values. The Leukemia cancer diagnosis data values are utilized in the classification process. Cancer risk level assessment is carried out in the classification process. The risk assessment model is divided into two phase. The feature selection or gene selection phase is applied to fetch the relevant gene elements. The classification phases performs the risk estimation using the Optimized Feature Selection based Support Vector Machine (OFS-SVM) technique. Criticality scores are estimated to perform the classification with high accuracy levels. The system can be enhanced with the following features. The classification system can be enhanced to support incremental data mining operations. The system can be improved to provide privacy preserved mining process. Stream based data classification process can be adapted in the classification system.

REFERENCES

[1] R. Braun, G. Leibon, S. Pauls, and D. Rockmore, “Partition Decoupling for Multi-Gene Analysis of Gene Expression Profiling Data,” BMC Bioinformatics, vol. 12, article 497, 2011, doi:10.1186/ 1471-2105-12-497.

[2] Z. Yu and H.-S. Wong, “Knowledge Based Cluster Ensemble for Cancer Discovery from Biomolecular Data,” IEEE Trans. Nano-Bioscience, vol. 10, no. 2, pp. 76-85, June 2011.

[3] Z. Yu and H.-S. Wong, “Class Discovery from Gene Expression Data Based on Perturbation and Cluster Ensemble,” IEEE Trans. NanoBioscience, vol. 8, no. 2, pp. 147-160, June 2009.

[4] Zhiwen Yu, Le Li, Hau-San Wong and Guoqiang Han, “SC3: Triple Spectral Clustering-Based Consensus Clustering Framework for Class Discovery from Cancer Gene Expression Profiles” IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 9, No. 6, November/December 2012.

[5] Arijit Ukil, Soma Bandyoapdhyay, Chetanya Puri and Arpan Pal, “IoT Healthcare Analytics: The Importance of Anomaly Detection”, IEEE 30th International Conference on Advanced Information Networking and Applications, 2016.

[6] C.Fernández-Llatas Martinez-Romero, Carvalho and Traver, “Challenges in Personalized Systems for Personal Health Care”, IEEE, 2016.

[7] Z. Yu and H.-S. Wong, “Knowledge Based Cluster Ensemble for Cancer Discovery from Biomolecular Data,” IEEE Trans. Nano-Bioscience, vol. 10, no. 2, pp. 76-85, June 2011.
[8] P. Qiu and S.K. Plevritis, “Simultaneous Class Discovery and Classification of Microarray Data Using Spectral Analysis,” J. Computational Biology, vol. 16, pp. 935-944, 2009.
[9] T.I. Simpson, J.D. Armstrong, and A.P. Jarman, “Merged Consensus Clustering to Assess and Improve Class Discovery with Microarray Data,” BMC Bioinformatics, vol. 11, article 590, 2010.
[10] N. Iam-on, T. Boongoen, and S. Garrett, “LCE: A Link-Based Cluster Ensemble Method for Improved Gene Expression Data Analysis,” Bioinformatics, vol. 26, no. 12, pp. 1513-1519, 2010.
[11] R. Avogadri and G. Valentini, “Fuzzy Ensemble Clustering Based on Random Projections for DNA Microarray Data Analysis,” Artificial Intelligence in Medicine, vol. 45, nos. 2/3, pp. 173-183, 2009.
[12] A. Bertoni and G. Valentini, “Discovering Multi-Level Structures in Bio-Molecular Data through the Bernstein Inequality,” BMC Bioinformatics, vol. 9 (Suppl 2), article S4, 2008.
[13] M. De Souto, I. Costa, D. De Araujo, T. Ludermir, and A. Schliep, “Clustering Cancer Gene Expression Data: A Comparative Study,” BMC Bioinformatics, vol. 9, article 497, 2008.