TIME SERIES ANALYSIS OF CANCER METASTASIS GENE EXPRESSION DATA USING PREDICTIVE CLUSTERING TREES (PCT)

Azariah Samuel. A[1] and Xavier Suresh. M[2]

[1] Research Scholar, Department of Bioinformatics, Sathyabama University, Chennai, Tamilnadu, India.
[2]Professor and Head, Department of Bioinformatics, Sathyabama University, Chennai, Tamilnadu, India.
Email: [1]azariahsamuel@gmail.com

Abstract

Time series microarray data analysis provides an invaluable insight into the genetic progression of metastasis. In this paper we analyzed time series microarray data of cancer metastasis. Predictive clustering using decision trees on time series data integrates clustering and prediction. We present an implementation of PCT on cancer metastasis gene expression data to obtain high level descriptions of clusters of gene expression samples in terms of clinical variables in an unparalleled serendipitous manner. We evaluate time series data from microarray experiments. Each data set records the change over time in the expression level of epithelial cells to facilitate dispersion by induction of Epithelial Mesenchymal Transition (EMT). Implementation of PCT with CLUS-TS on the cancer metastasis gene expression data enables us to understand the early event during tumor metastasis. Summing up, we applied predictive clustering trees to the cancer metastasis gene expression data, where the goal was to obtain integrated high-level descriptions of clusters of gene expression samples in terms of clinical variables. Our preliminary results hint at the potential to understand the molecular pathogenesis of tumor metastasis through predictive clustering using decision trees approach.

Key words: Predictive Clustering Trees (PCT), CLUS-TS, Cancer Metastasis and Time Series Microarray Data.

I. INTRODUCTION

Metastasis or metastatic disease is the spread of a disease from one organ or part to another non-adjacent organ or part [1]. Metastatic tumors are very common in the late stages of cancer. The spread of metastases may occur via the blood or the lymphatic system or through both routes. The most common places for the metastases to occur are the lungs, liver, brain, and the bones [2]. Metastasis is a complex series of steps in which cancer cells leave the original tumor site and migrate to other parts of the body via the bloodstream or the lymphatic system. To do so, malignant cells break away from the primary tumor and attach to and degrade proteins that make up the surrounding extracellular matrix (ECM), which separates the tumor from adjoining tissue. By degrading these proteins, cancer cells are able to breach the ECM and escape [3].

A time series is a sequence of data points, measured typically at successive times spaced at uniform time intervals. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series data have a natural temporal ordering. This makes time series analysis distinct from other common data analysis problems, in which there is no natural ordering of the observations. A time series model will generally reflect the fact that observations close together in time will be more closely related than observations further apart [4].

Predictive clustering is a general framework that combines clustering and prediction [5]. Predictive clustering partitions a given data set into a set of clusters such that the instances in a given cluster are similar to each other and dissimilar to the instances in other clusters. In this sense, predictive clustering is identical to regular clustering [6]. The difference is that predictive clustering associates a predictive model to each cluster. This model assigns instances to clusters and provides predictions for new instances. So far, decision trees [5, 7] and rule sets [8] have been used in the context of predictive clustering. This paper investigates how predictive clustering can be applied to cluster time series [9]. A time series is an ordered sequence of measurements of a continuous variable that changes over time.

Description of the data

The data were obtained from the curated dataset browser of NCBI's Gene Expression Omnibus, a public functional genomics data repository supporting MIAME compliant data submissions. The obtained data is a time series data of Transforming growth factor (TGF) *beta* treatment of A549 lung adenocarcinoma cell line of Homo sapiens. The experiment profiled temporal gene expression changes during TGF – beta - induced Epithelial - Mesenchymal Transition (EMT). During EMT cancer cells lose their epithelial specific proteins and gain mesenchymal proteins to acquire migratory and invasive phenotype essential for metastasis. EMT is a developmental process that facilitates the dispersion of cells. A similar process is reactivated in cancer cells as an early event during tumor metastasis. Samples were assayed using Affymetrix HG_U133_plus_2 arrays with 54675 probe-sets, using standard techniques [10].

Clustering Time series Data

A new algorithm called Clus-TS (Clustering-Time Series) that constructs trees instantiates the general PCT induction algorithm to the task of time series clustering [5]. We evaluate Clus-TS on time series data from microarray experiments [14]. Each data set records the change over time in the expression level of Transforming growth factor (TGF) beta treatment of A549 lung adenocarcinoma cell line of Homo sapiens. A lot of work has been previously done by clustering this type of short time series gene expression data by J. Ernst et al [11]. We use a different distance measures that mainly take the shape of the time series into account and to construct clusters that can be explained by a given set of features. Besides the time series, various other data about each gene are available. Here, we consider motifs and terms from the Gene Ontology (GO) [12]. The motifs are subsequences that occur in the amino acid sequence of many genes. The motivation for using motifs as features is due to Curk et al. [13], who use motifs in a similar analysis. The motifs or GO terms appear in the internal nodes of the PCT and provide a symbolic description of the clusters. This is related to itemset constrained clustering [15], which clusters vectors of numeric values and constrains each cluster by means of an item set. Researches on inductive databases (IDBs) [16, 17] have focused on local models (i.e., models that apply to only a subset of the examples), such as frequent item sets and association rules. Clus-TS is part of a larger project [18, 7, 19] were the goal is to investigate how IDBs can be extended to global models, such as decision trees (for prediction) and mixture models (for clustering). Predictive clustering has been argued to provide a general framework unifying clustering and prediction, two of the most basic data mining tasks, and is therefore an excellent starting point for extending IDBs to global models [19]. Extending PCTs to time series clustering is a system that is applicable to clustering and prediction in many application domains, including bioinformatics [14].

II. PREDICTIVE CLUSTERING TREES

A. Prediction, Clustering, and Predictive Clustering Trees.

Predictive modeling aims at constructing models that can predict a target (T) property of an object from a description (D) of the object. Predictive models are learned from sets of examples, where each example has the form (D, T). Clustering [20], on the other hand, is concerned with grouping objects into subsets of objects (called clusters) that are similar with respect to their description D. Predictive clustering [5] combines elements from both prediction and clustering. The predictive model assigns new instances to clusters based on their description D and provides a prediction for the target property T. A well-known type of model that can be used to this end is a decision tree [21]. A decision tree that is used for predictive clustering is called a predictive clustering tree. Each node of a PCT represents a cluster. The conjunction of conditions on the path from the root to that node gives a description of the cluster. Essentially, each cluster has a symbolic description in the form of a rule (IF conjunctions of conditions THEN cluster) [22], while a tree structure represents the hierarchy of clusters. Clusters that are not on the same branch of a tree do not overlap. The description D of a gene consists of GO terms with which the gene is annotated, and the target property T is the time series recorded for the gene. In general, we could include both D and T in the distance measure. We are, however, most interested in the time series part. Therefore, we define the distance measure only on T. The resulting PCT represents a clustering that is homogeneous with respect to T and the nodes of the tree provide a symbolic description of the clusters. A PCT can also be used for prediction: use the tree to assign a new instance to a leaf and take the centroid of the corresponding cluster as prediction [14].

B. Building Predictive Clustering Trees

The generic induction algorithm for PCTs [5] is a variant of the standard greedy recursive top-down decision tree induction algorithm [21]. It takes as input a set of instances I (genes described by motifs or GO terms and their associated time series). The procedure BestTest [5] searches for the best acceptable test (motif or GO term) that can be put in a node. If such a test t* can be found then the algorithm creates a new internal node labeled t* and calls itself recursively to construct a subtree for each cluster in the partition P* induced by t* on the instances. If no acceptable test can be found, then the algorithm creates a leaf, and the recursion terminates. (The procedure Acceptable defines the stopping criterion of the algorithm, e.g., specifying maximum tree depth or a minimum number of instances in each leaf). Till here, the algorithm is identical to a standard decision tree learner. The main difference is in the heuristic that is used for selecting the tests. For PCTs, this heuristic is the reduction in variance (weighted by cluster size). Maximizing variance reduction maximizes cluster homogeneity [14]. An implementation of the PCT induction algorithm is available in the Clus system, which can be obtained at http://www.cs.kuleuven.be/~dtai/clus [14].

C. PCTs for Time Series Clustering

In this section, we discuss a number of distance measures for time series, which will be used in the definition of cluster variance.

Dynamic Time Warping (DTW) [23] can capture a non-linear distortion along the time axis.

Correlation [14]

The correlation coefficient r(X,Y) between two time series X and Y is calculated as $r(X, Y) = E[(X - E[X]) \bullet (Y - E[Y])]/E[(X - E[X])^2]\bullet E[(Y - E[Y])^2]$, where E[V] denotes expectation (i.e., mean value) of V . r(X,Y) measures the degree of linear dependence between X and Y. It has the following intuitive meaning in terms of the shapes of X and Y: r close to 1 means that the shapes are similar. If there is a linear relation between X and Y then the time series is identical but might have a different scale or baseline. r close to -1 means that X and Y have "mirrored" shapes, and r close to 0 means that the shapes are unrelated (and consequently dissimilar). Based on this intuitive interpretation, we can define the distance between two time series as $d_r(X, Y) = \text{sqrt } (0.5 \bullet (1 - r(X, Y))$.

Qualitative Distance

A third distance measure is the qualitative distance proposed by Todorovski et al. [24]. It is based on a qualitative comparison of the shape of the time series.

III. TIME SERIES ANALYSIS OF METASTATIC CARCINOMA GENE EXPRESSION DATA USING PREDICTIVE CLUSTERING TREES

Itemset Constrained clustering

Gene expression analysis of time series data is a new arena for making meaningful results in microarray informatics. Clustering genes by their time expression pattern makes sense because genes which are co-regulated or have a similar function, under certain conditions, will have a similar temporal functional profile. Instead of simply clustering the expression time series with hierarchy and later on elucidating the characteristics of the obtained clusters as done in e.g., [11], we perform constrained clustering with PCTs. This yields the clusters and symbolic descriptions of the clusters in one step. We use the data from a study conducted by Santor et al. [10]. The purpose of the study is to explore the changes in expression levels of Transforming growth factor (TGF) beta treatment of A549 lung adenocarcinoma cell line of Homo sapiens. There were 9,921 new probe sets representing approximately 6,500 new genes.

The steps for this analysis are:

- Find frequent item sets from the cancer metastasis data records (attributes);
- Use them as features;
- Create PCT stubs with features as constraints.

The Data Mining

It consists of two steps. In the first step, we use a local pattern mining algorithm to construct patterns based on the description of the genes. In a second step, we use these local patterns as features to construct PCTs. We use two types of features: motifs and GO terms [12]. For the first set of features, we mine frequent subsequences (motifs) occurring in the DNA sequences of the genes, which we obtain from the GEO repository. We use the constraint based mining algorithm FAVST [25, 26]. FAVST supports three types of constraints: minimum frequency, and minimum and maximum motif length. We query FAVST for sequences that appear in 25% of the genes and consist of at least 8 nucleotides. In this way, we obtain 387 motifs ranging from 8 to 10 nucleotides. These motifs are passed to Clus-TS to build PCTs with the motifs in the internal nodes. In the second set of features, each feature is a GO term. We obtain the GO term annotations for each gene from the Gene Ontology [12]. The GO terms are structured in a hierarchy. To limit the number of features, we set a minimum frequency threshold: each GO term must appear for at least 50 of the genes.

Determining significant Gene Expression profiles

To determine significant gene expression profiles on the time course data, we consider genes or set of genes that might serve TGF-beta-induced epithelial-mesenchymal transition (EMT) at various time course of the experiment. We used Clus-TS in a beam search mode in order to see which, the best multi-objective trees that can be constructed. We found out that the top 216 trees have the same value for the intra-cluster variance. Each of these trees had three genes in its internal nodes. We found out that there are 216 trees have the same value for the intra-cluster variance. Each of these trees had three genes in its internal nodes.

IV. RESULTS - PREDICITVE CLUSTERING TREES

There were 216 models of multi objective classification trees, which we consider the best models of the cancer metastasis gene expression data. The probes which were part of these models were identifies and ranked according to the number of appearances in the model. In this way we identified, 41 probes. To see if any probes (genes) are identified as significant in the biological view point, we searched their function from the gene ontology database.

Among the 41 probes, 22 of them possess little or no biological function relevant to the metastatic potential and cannot be correlated to a known gene of biological function. Probes "1566324_a_at", "218559_s_at", 211607_x_at", "211421_s_at", "210511_s_at", 205051_s_at", "205249_at", 1555997_s_at", "204259_at" are involved in processes related to oncogenic metastasis. Probes “206026_s_at", "215913_s_at", "209789_at", "212489_at" are involved in processes related to transcriptional activator activity, transcription factor activity, dTDP, dTTP biosynthesis and DNA metabolism. Various studies have previously documented disturbed transcriptional regulation is a driving force in the oncogenic metastatic progression. Therefore our finding that expression of these transcriptions related genes are different and suggests that the transcription is disturbed in cancer metastasis. Also, probes "210118_s_at", "235342_at", "225369_at", "22314_x_at", "237411_at", "214404_x_a" are involved in processes related to immunomodulation, intercellular signal cascading, protein functioning, synthesis and degradation, proteolysis and pepdidase activity, protein modification, protein binding, protein amino acid glycosylation, protein biosynthesis, translational initiation or regulation of translation.

V. CONCLUSION AND FUTURE WORK

In this study a lot of effort was made in clustering genes using predictive clustering trees approach to produce biologically meaningful results. As a part of analysis, microarray data of cancer metastasis was made to demonstrate the use of time series PCTs for the expression of changes during TGF – beta - induced Epithelial - Mesenchymal Transition (EMT) in various time intervals and to identify significant gene expression profiles. Some of the genes identified and listed above could prove useful as biomarkers for cancer metastasis. However, validation of the results by repeating the analysis on individual cancer metastasis data sets and comparing results from different types of analysis and previous studies are essential, since each of the secondary tumors associated with cancer metastasis expresses a different phenotype. Further work and deeper biological insights are needed in order to be able to make any kind of a significant claim and true validation of the results.

ACKNOWLEDGEMENT

I am thankful to my supervisor, Dr. M. Xavier Suresh, Professor and head, Department of Bioinformatics, Sathyabama University, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject.

REFERENCES

[1] Klein CA (September 2008). "Cancer, The metastasis cascade". Science 321 (5897): 1785–7.

[2] Metastatic Cancer: Questions and Answers. National Cancer Institute.

[3] Yoshida BA, Sokoloff MM, Welch DR, Rinker-Schaeffer CW (Nov 2000). Metastasis-suppressor genes: a review and perspective on an emerging field. J Natl Cancer Inst. 92 (21): 1717–30.

[4] Gershenfeld.N. (1999).The nature of mathematical modeling. p. 205-08.

[5] H. Blockeel, L. De Raedt, and J. Ramon. Top-down induction of clustering trees. In 15th Int'l Conf. on Machine Learning, pages 55–63, 1998.

[6] L. Kaufman and P.J. Rousseeuw, editors. Finding groups in data: An introduction to cluster analysis. John Wiley & Sons, 1990.

[7] J. Struyf and S. D¡zeroski. Constraint based induction of multi-objective regression trees. In 4th Int'l Workshop on Knowledge Discovery in Inductive Databases: Revised Selected and Invited Papers, volume 3933 of LNCS, pages 222–233. Springer, 2006.

[8] B. ¡Zenko, S. D¡zeroski, and J. Struyf. Learning predictive clustering rules. In 4th Int'l Workshop on Knowledge Discovery in Inductive Databases: Revised Selected and Invited Papers, volume 3933 of LNCS, pages 234–250. Springer, 2005.

[9] T.W. Liao. Clustering of time series data, a survey. Pattern Recognition, 38: 1857–1874, 2005.

[10] Sartor MA, Mahavisno V, Keshamouni VG, Cavalcoli J. ConceptGen: a gene set enrichment and gene set relation mapping tool. Bioinformatics 2010 Feb 15; 26(4): 456-63.

[11] J. Ernst, Nau G.J., and Bar-Joseph Z. Clustering short time series gene expression data. Bioinformatics, 21(Suppl. 1):159–168, 2005.

[12] M. Ashburner et al. Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nature Genet., 25(1): 25–29, 2000.

[13] T. Curk, Zupan B., Petrovi¡c U and Shaulsky G. Ra¡cunalni¡sko odkrivanje mehanizmov uravnavanja istra¡zanja genov. In Prvo sre¡canje slovenskih bioinformatikov, pages 56–58, 2005.

[14] Sašo D_eroski, Valentin Gjorgjioski, Ivica Slavkov and Jan Struyf. Analysis of Time Series Data with Predictive Clustering Trees. Knowledge Discovery in Inductive Databases Lecture Notes in Computer Science, 2007, Volume 4747/2007, 63-80.

[15] J. Sese, Y. Kurokawa, M. Monden, K. Kato, and S. Morishita. Constrained clusters of gene expression profiles with pathological features. Bioinformatics, 20: 3137– 3145, 2004.

[16] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of the ACM, 39(11):58–64, 1996.

[17] L. De Raedt. A perspective on inductive databases. SIGKDD Explorations, 4(2): 69–77, 2002.

[18] E. Fromont and H. Blockeel. Integrating decision tree learning into inductive databases. In 5th Int'l Workshop on Knowledge Discovery in Inductive Databases, pages 59–70, 2006.

[19] B. Zenko, S. D¡zeroski, and J. Struyf. Learning predictive clustering rules. In 4th Int'l Workshop on Knowledge Discovery in Inductive Databases: Revised Selected and Invited Papers, volume 3933 of LNCS, pages 234–250. Springer, 2005.

[20] L. Kaufman and P.J. Rousseeuw, editors. Finding groups in data: An introduction to cluster analysis. John Wiley & Sons, 1990.

[21] J.R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann series in Machine Learning. Morgan Kaufmann, 1993.

[22] R.S. Michalski and R.E. Stepp. Learning from observation: conceptual clustering. In Machine Learning: an Artificial Intelligence Approach, volume 1. Tioga Publishing Company, 1983.

[23] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spokenword recognition. In IEEE Transaction on Acoustics, Speech, and Signal Processing, volume ASSP-26 of LNAI, pages 43–49, 1978.

[24] L. Todorovski, B. Cestnik, M. Kline, N. Lavra¡c, and S. D¡zeroski. Qualitative clustering of short time-series: A case study of firms reputation data. In ECML/PKDD'02 Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning, pages 141–149, 2002.

[25] S.D. Lee and L. De Raedt. An efficient algorithm for mining string data-bases under constraints. In 3th Int'l Workshop on Knowledge Discovery in Inductive Databases: Revised Selected and Invited Papers, volume 3377 of LNCS, pages 108–129. Springer, 2004.

[26] I. Mitasiunaite and J-F. Boulicaut. Looking for monotonicity properties of a similarity constraint on sequences. In ACM Symposium of Applied Computing SAC'2006, Special Track on Data Mining, pages 546–552. ACM Press, 2006.

Mr. A. Azariah Samuel has completed his Master of Technology in the Faculty of Bioinformatics, Sathyabama University, in the year 2009 and currently works as Assistant Professor in the Department of Bioinformatics, VMKV Engineering College, Vinayaka Missions University, Salem. He is a Ph.D research scholar of Sathyabama University, under the supervision of Dr. M. Xavier Suresh, Professor and Head, Department of Bioinformatics, Sathyabama University, Chennai.