Analysis of Huntington's disease gene expression profiles using Constrained Clustering

Ivica Slavkov [1], Sašo Džeroski [1], Borut Peterlin [2] , Luca Lovrecic [3]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[2] Department of Gynaecology and Obstetrics, University Medical Centre Ljubjana, Ljubljana , Slovenia
[3] Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02128, USA

Abstract

In this paper we looked at the possibility of analyzing gene expression data by using constrained clustering. The clustering was used as a way to relate Huntington's disease patient pathological data with its corresponding microarray data, and it was performed by using a generic system for constructing decision trees. This resulted in genes whose expression level could possibly serve as an indicator of the progress of the disease.

Introduction

When analyzing gene expression data (ex. microarray) in order to extract meaningful patterns it would be useful to relate somehow the gene expression levels to the patients' pathological data. The **I**temset-**C**onstrained clustering developed by Morishita et al. [2] is one way of doing this. However, **IC**-clustering as such, reveals which patient pathological attributes create biggest difference between the gene expression profiles. For example, in the results obtained in [2] , the cluster labeled with "Tumor" was at the top of the list of informative clusters, which is expected, but also a cluster "Tumor+Man" (which was overlooked by k-means clustering) points out that different gene expression profiles exist in male and female tumor patients. In this paper a modified approach in discovering informative clusters is presented. Constrained clustering is carried out on Huntington's disease patient records and microarray data, by using the gene expression levels as constraints. As a result we obtain informative clusters labeled by certain genes.

Huntington's Disease (HD)

Huntington's disease is an autosomal dominant neurodegenerative disorder characterized by progressive motor impairment, cognitive decline, and various psychiatric symptoms, with the typical age of onset in the third to the fifth decade. It is caused by glutamine repeats in ubiquitously distributed huntingtin protein. Recent studies have shown that mutant huntingtin interferes with the function of widely expressed transcription factors, suggesting that gene expression may be altered in a variety of tissues, including peripheral blood. That is why in [3] it has been postulated that microarray data obtained from peripheral blood samples for HD patients could be used to analyze the changes of gene expression levels which HD causes.

In this paper the dataset which is used for analysis of HD gene expression profiles is one of the datasets that are used in [3]. The microarray data was obtained from the Amersham Biosciences platform. It contained information about the gene expression levels of 20.383 different genes, for HD symptomatic (11) and presymptomatic (5) patients. On the other hand, patient pathological data included nominal attributes like: Symptomatic/Presymptomatic, Age>40, Sex {Male, Female}, Stage {Early, Mid, Late} and TFC>10 (Total Functional Capability).

Our purpose was by using constrained clustering to relate this microarray data to the patients' pathological data. We considered that it would reveal informative clusters, i.e., genes (and their corresponding expression levels) which could be regarded as ones which can be used to distinguish between symptomatic and presymptomatic patients and reveal information about the progress (stage) of the disease.

Constrained Clustering using Clus

In order to perform the constrained clustering we use a generic system for constructing decision trees- **Clus**. It can be used to create classification trees for predicting symbolic attributes and regression trees for predicting numerical values. In some cases it is useful to predict several attributes at the same time, so multi-objective decision trees can also be constructed. Clus can generate so-called Predictive Clustering Trees (PCTs) [1], using a beam search algorithm, which actually are decision trees that are used for hierarchical clustering purposes. The data that we have consists of patient records which contain different information about the HD patients and we also have their corresponding gene expression levels as microarray data. Considering this, we use Clus to construct the so-called PCT stubs, i.e., decision trees with only one test node. The test is selected from the gene expression levels data and it imposes a binary split (clustering) on the samples. Then the intra-cluster variance is calculated for the PCT stub, by using the patient pathological data. We search for the best *N* PCT stubs, by using the beam search algorithm with beam width *N*. Each PCT stub is evaluated by its intra-cluster variance and if it is found to be suitable, then it is put in the current search beam of size N.

In the end we get a list of the N best PCTs (clusters) which are sorted by the intra-cluster variance.

Results

In this section we present the results obtained with constrained clustering on the previously described HD patient/microarray data. They are presented in the fowolling table.

Table 1 Genes and their corresponding expression levels obtained by constrained clustering

Cluster Rank	Probe name > expression level	Intra-cluster variance
1	1098576.1_PROBE1 > 110,662	10.694
2	NM_005679.1_PROBE1 > 3,506	10.694
3	321334.1_PROBE1 > 0,095	10.694
4	AF038190_PROBE1 > 0,152	10.694
5	NM_013366.2_PROBE1 > 0,763	10.694
6	NM_021170.1_PROBE1 > 0,965	10.694
7	NM_004662.1_PROBE1 > 0,142	10.694
8	105763.3_PROBE1 > 2,928	10.694
9	1137177.1_PROBE1 > 0,469	10.694
...	...	...
70	1118860.1_PROBE1 > 9,996	10.694

In order to do a further selection of the top 70 genes, which had the same intra-cluster variance, we filtered them by using various tests. First a Student *t* test was performed and only genes with $P < 0.\ 0005$ were considered. Also genes which had expression ratio of average Presymptomatic/average Symptomatic >1.8 or <0.6 as cut-off values were selected. The genes with low expression levels have lower copy numbers of mRNA and are therefore more susceptible to technical noise. That is why the last filter produced only genes in which at least one sample has value greater then 1.

Table 2

Gene N°	Selected genes
1	1098576.1_PROBE1 > 110,662
2	1137177.1_PROBE1 > 0,469
3	AB002323_PROBE1 > 17,718
4	NM_002248.2_PROBE1 > 6,333
5	1500797.5_PROBE1 > 4,533
6	AA682448_PROBE1 > 8,36
7	1501054.6_PROBE1 > 23,306
8	NM_014059.1_PROBE1 > 7,933
9	1013124.1_PROBE1 > 1,243
10	2814863CB1_PROBE1 > 38,966
11	1452625.4_PROBE1 > 57,164
12	481822.1_PROBE1 > 6,592
13	1133814.1_PROBE1 > 1,31
14	1113924.1_PROBE1 > 0,663
15	NM_018169.1_PROBE1 > 70,09
16	1118860.1_PROBE1 > 9,996

At the end 16 genes were selected and they are presented in table 2. Here it is interesting to note that after examining the clusters that are described by these 16 genes, all of them had grouped together patients which were presymptomatic, with those that were symptomatic but in the early stage of the development of the disease. This implies that presymptomatic patients have similar gene expression levels as those of symptomatic but in the early stage of HD. This means that big changes in gene expression levels in the blood samples of HD patients can be detected in advanced stages of the disease.

Furthermore, we analyzed the HD data also by performing constrained clustering, but this time we used the patient records as constraints. The clusters that we obtained revealed which patient pathological features had biggest difference in the gene expression levels. The results are presented in table 3.

Table 3

Cluster Rank	Patient pathological features	Intra-cluster variance
1	Symptomatic = 0	723.694
2	StageEarly = 0	928.005
3	Symptomatic_Age>40 = 0	993.234
4	TFC>=10 = 0	1035.076
5	SexMale_StageEarly = 0	1088.87
6	SexMale_TFC>=10 = 0	1090.29
7	Symptomatic_Age>40_TFC >=10 = 0	1093.576
8	Symptomatic_SexMale = 0	1094.41
9	StageMid = 0	1095.02
...	...	...

As expected, the top ranked cluster is the "Symptomatic" cluster. The cluster ranked as second is "StageEarly", which supports the previous claim that bigger difference of gene expression levels exist in patients that are in early stage of the disease and those that are in the mid or late.

Conclusion and further work

The purpose of this paper was to demonstrate how constrained clustering can be used in determining gene expression profiles which could serve as indicators about different aspects of a disease. However, in order to confirm the validity of the results obtained in analyzing of the Huntington's disease data, further testing is needed on more datasets and also work has to be carried out concerning feature selection in order to reduce the ratio of number of genes/sample size.

References

1. H. Blockeel, L. De Raedt, and J. Ramon. :Top-down induction of clustering trees. In Proceedings of the 15th International Conference on Machine Learning, pages 55–63, 1998.
2. Sese Jun, Yukinori Kurokawa, Kikuya Kato, Morito Monden and Shinichi Morishita.: Constrained Clusters of Gene Expression Profiles with Pathological Features. Bioinformatics vol. 20 issue 17 Oxford University Press 2004.
3. Borovecki et al. : Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. In Proceedings of the National Academy of Sciences of the USA, August 2 2002, vol 102., no 31, p 11023-11028