International Journal of Advanced Research in

Computer Science Engineering and Information Technology

Volume: 2 Issue: 2 08-Mar-2014,ISSN_NO: 2321-3337

Mining signature of heterogeneous event sequences using beta divergence to develop highly robust data

D.Gayathri, Information Technology, Panimalar Engineering College,India.
F.Evelin Rosy, Information Technology, Panimalar Engineering College,India.
D.Vani, Information Technology, Panimalar Engineering College,India.

Abstract- *Large collections of electronic clinical records today provide us with a vast source of information on edical practice. However, the utilization of those data for exploratory analysis to support clinical decisions is still limited. Extrmacting useful patterns from such data is particularly challenging because it is longitudinal, sparse and heterogeneous. In this paper, we propose a Nonnegative Matrix Factorization based framework using a convolutional approach for open-ended temporal pattern discovery over large collections of clinical records. We call the method One-Sided Convolutional NMF. Our framework can mine common as well as individual shift-invariant temporal patterns from heterogeneous events over different patient groups, and handle sparsity as well as scalability problems well. Furthermore, we use an event matrix based representation that can encode quantitatively all key temporal concepts including order, concurrency and synchronicity. We derive efficient multiplicative update rules for OSC-NMF, and also prove theoretically its convergence. Finally, the experimental results on both synthetic and real world electronic patient data are presented to demonstrate the effectiveness of the proposed method.*

Keywords-**Temporal pattern, Non negative matrix, Synthetic data**

1.INTRODUCTION

Extracting clinical records for analysis is salient as it helps in predicting of various diseases in the medical field. Other Patient's record is used for diagnosing disease for the newly registered patient. Latent signature mining facilitates decision support for patient diagnosis, prognosis and management. Event knowledge Representation is introduced in order to enable easy and efficient understanding of patient's detail. Temporal Event Matrix Representation enables us to represent the patient's report values in a matrix format. One dimension of the matrix should represent the event and other dimension should represent the time of the corresponding event. Cancer is one of the most prevalent disease in of cancer using clinical records of different patients is highly challenging.

1.1 Breast Cancer-Introduction

Breast cancer is a malignant tumor that starts in the cells of the breast. A malignant tumor is a group of cancer cells that can grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body. The disease occurs almost entirely in women, but men can get it, too. Most breast cancers begin in the cells that line the ducts (*ductal* cancers). Some begin in the cells that line the lobules (*lobular* cancers), while a small number start in other tissues

1.2 Lymph system of the breast

The lymph system is important to understand because it is one way breast cancers can spread. This system has several parts.

Lymph nodes are small, bean-shaped collections of immune system cells (cells that are important in fighting infections) that are connected by lymphatic vessels. Lymphatic vessels are like small veins, except that they

carry a clear fluid called *lymph* (instead of blood) away from the breast. Lymph contains tissue fluid and waste products, as well as immune system cells. Breast cancer cells can enter lymphatic vessels and begin to grow in lymph nodes.

Most lymphatic vessels in the breast connect to lymph nodes under the arm (*axillary nodes*). Some lymphatic vessels connect to lymph nodes inside the chest (*internal mammary nodes*) and those either above or below the collarbone (*supraclavicular* or *infraclavicular nodes*)

2. RELATED WORK

2.1 Data mining to mine heterogeneous event data:

Today, the size of the population, the amount of electronic data gathered, along with globalization and the speed of disease outbreaks make it almost impossible to accomplish what the pioneers did. This is where data mining becomes useful to healthcare.

Data mining as utilization of statistical techniques within the knowledge discovery process. When medical institutions apply data mining on their existing data, they can discover new, useful and potentially life-saving knowledge that otherwise would have remained inert in their databases.

“Data mining And decision support methods, including novel visualization methods, can lead to better performance in decision-making. Data mining allows organizations and institutions to get more out of existing data at minimal extra cost. By using data mining and visualization, medical experts could find patterns and anomalies better than just looking at a set of tabulated data.

2.2 Prediction of cancer using data mining techniques :

Prediction of cancer can be done using three data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. This paper proposes the idea of predicting cancer using the clinical records of patients detail. The patients test report field values are matched against each other in order to accurately find what type of cancer the patient poses.

3.SYSTEM ANALYSIS :

3.1 EXISTING SYSTEM

In the Existing system , prediction of disease using large amount of clinical records was done using simple data mining techniques .There was no filtering criteria’s used for further accurate extraction of records. The records retrieved using a simple data mining techniques like field value matching resulted in the retrieval of large collections of data and hence it resulted in time consumption for disease prediction .

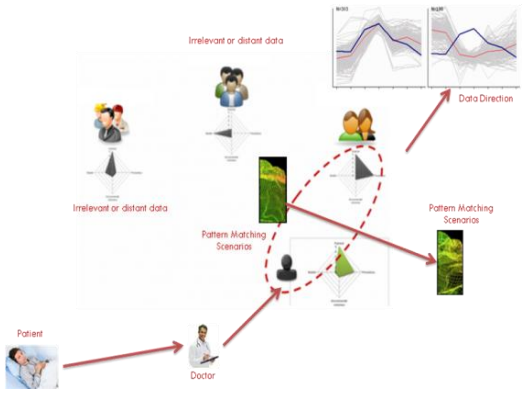
3.2 PROPOSED SYSTEM

In the proposed system, filtering criterias are added to simple data mining techniques. This incorporation of criteria’s resulted in further extraction of retrieved records. When the extraction is extended, the records resulted finally were minimal and hence the task of predictionality was made easier by using limited time period.

4. PREDICTION USING HIGHLY ROBUST DATA PATTERN

To identify an interesting real time problem of one’s health by means of Health Care Data which is based on the history of one patient by matching the health history of other patients

Architecture



A temporal signature is a detectable phenomenon which defines an object's position in time.
Pattern matching scenarios is used to match the symptoms of one patient with another.

The patient reports about the symptoms to the doctor.The patient history needs to be fed into the system. An appropriate information about the relevant data will be matched from the system. System is provided with Factor based temporal matching algorithm to match with relevant contents from historical databases.The graph is generated by comparing the datas of the patient.

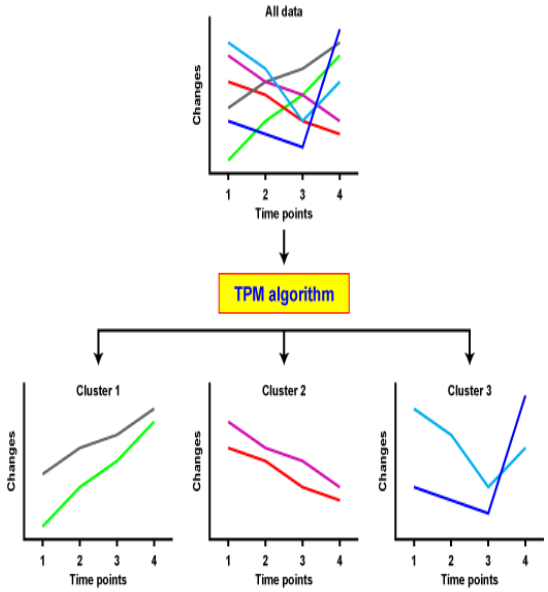
4.1 Temporal pattern generation for a given patient :

The temporal pattern discovery framework addresses three issues simultaneously: statistical shrinkage reduces the risk of highlighting spurious associations and allows for effective large-scale screening , a statistical graphical approach to summarising and visualising event history data facilitates open ended exploratory analysis without prior restrictions on the types of temporal patterns considered, and a comparison of the time period of interest to a control period prior to the drug prescription allows true temporal association to be identified.
Temporal pattern generation is initially done for the registered patient.
The database contains list of patterns of patients with similar disease and it is grouped accordingly. Based on the values formulated , pattern search and grouping is done accordingly to produce a robust matched patient's data . temporal pattern can be represented in different formats like graphical representation . object representations like triangle, circle , rectangle , square etc can also be used to represent the temporal pattern .
Different shapes are adopted for security purpose .

Temporal pattern generation method

Required:G(Given patient),S(Set of patients record),T
Method:Generate pattern for G
for T = 1 to S do
matching field found with G
if reached S
break
endif
endfor
return pattern generated

Temporal pattern matching



4.2 Mining signatures from multiple event sequences using beta – divergence:

Temporal data in the form of event sequences are generated . each event sequences contain different field values . temporal pattern associations in multiple event sequences is a challenging task . it can be resolved by using beta – divergence method.

Temporal data in the form of event sequences are generated . each event sequences contain different field values . temporal pattern associations in multiple event sequences is a challenging task . it can be resolved by using beta – divergence method.

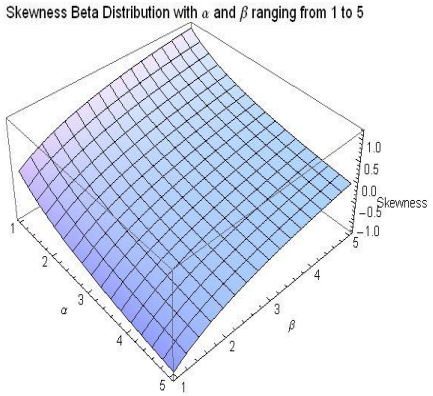
beta -divergence is a family of cost functions parameterized by a single shape parameter beta that takes the Euclidean distance, the Kullback-Kibler distance and the Itakura-Saito divergence as special cases(beta=2,1,0 respectively).

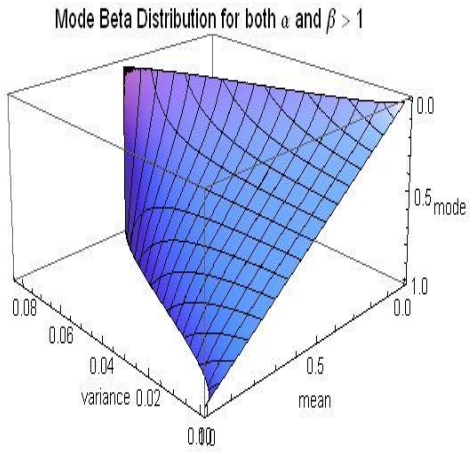
It makes use of aximization-equation(ME) algorithm which produces updates that move along constant level sets of, the auxiliary function and leads to large steps then majorization-minimization(MM). this is akin to over relaxation and is show experimentally, to produce faster convergence

betw-diverge can be designed as

d beta(x/y) = {1/beta(beta-1)(x pow beta+(beta-1)y pow beta-beta x y
pow beta-1 beta E R\{0,1}x logx/y-x+y beta=1
x/y-logx/y-1 beta=0}

the beta-divergence can be shown continuous in beta by using the identity lim beta->0(x pow beta- y pow beta)/beta=log (x/y). the limit cases beta=0 and beta=1 correspond to IS and KL divergences





OSC-NMF method

Require:X,F,G,r,T,BETA,LAMDA
Ensure:F>=0; G>=0
Initalize F,G
For i=1 to T do
Update F
Update G
If converged then
break
endif
endfor
return R={W,H}

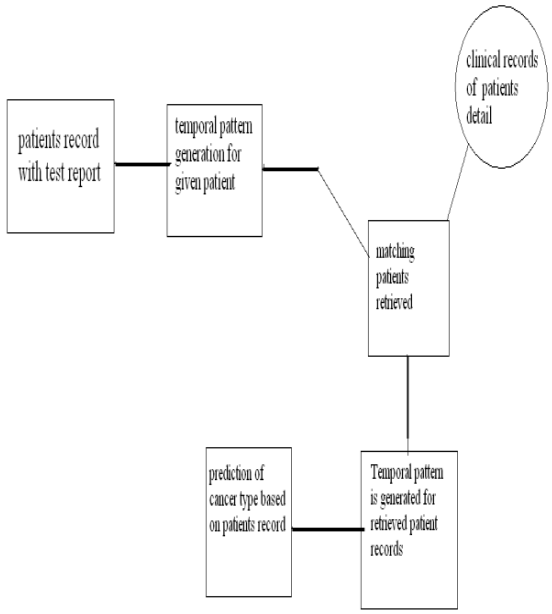
4.3 Consolidating data and generating synthetic dataset

All the retrieved patterns ,ie the pattern got from cardiogram , cancer and bupa test is consolidated as a single format .Consolidation is important to get accurately matched patients detail for all the 3 test then for the consolidated data pattern synthetic dataset is generated.Synthetic data is an involved process of data anonymization; that is to say that synthetic data is a subset of anonymized data from the anonymized data certain fields information are represented in graphical format .consolidating data of various test taken helps doctor to verify the patients record matched for each of the test taken.

4.4 Analysis of breast cancer types

Finding the cancer can be done by taking the appropriate test based on the symptoms. Each and every cancer has its own classification. In order to predict it , separate test need to be taken. In our paper, we say that there is no need to take separate test to find cancer type. By using the clinical records of patient , cancer classification can be easily predicted. More accurate prediction is possible and once comparison is done , treatment can be started based on the treatment adopted for other patients. In our paper, treatment method to be carried out can be known based on the matched patients record .

Synthetic data is used in a variety of fields as a filter for information that would otherwise compromise the confidentiality of particular aspects of the data.Synthetic data are generated to meet specific needs or certain conditions that may not be found in the original, real data. This can be useful when designing any type of system because the synthetic data can be used as a representation of theoritical value of fields .



5. EXPERIMENTAL RESULTS

5.1 Predicition of cancer using health care data

Blood cancer can be specifically investigated for their capacity of predicting venous thromboembolism(VTE) during the course of disease. Parameters of blood count analysis (elevated leukocyte and platelet-count, decreased haemoglobin) have turned out to be useful in risk prediction. Associations between elevated levels and future VTE have been found for D-Dimer, prothrombin fragment 1+2 and soluble P-selectin and also for clotting factor VIII and the thrombin generation potential. The results for tissue factor (TF)-bearing microparticles are heterogeneous, an association with occurrence of VTE in pancreatic cancer might be present. whereas in other cancer entities, such as glioblastoma, colorectal or gastric carcinoma this could not be confirmed.

Example :

Consider a clinical record of 7 patients taken as sample .Each patient possess different breast cancer .

Patient Id	Clump Thickness	Marginal Adhesion	Cancer Type
1	2	10	Ductal carcinoma
2	2	1	Invasive ductal
3	1	5	Triple negative
4	3	1	Metastatic breast cancer
5	1	1	Invasive ductal

The given patient field values are as follows ,

Patient id	Clump thickness	Marginal Adhesion	Cancer Type
6	2	4	Triple Negative

For the given patient the type of breast cancer can be accurately predicted based on the comparison of the field values .

4.2 Prediction of various other disease

In humans, *Mycobacterium tuberculosis* persists for long periods in a clinically latent state, creating a huge reservoir of 'silent' tuberculosis (TB) (roughly one-third of the global population) from which new cases continually arise. A prognostic marker for active TB would enable targeted treatment of the small fraction of infected individuals who are most at risk of developing contagious TB, contributing greatly to TB control efforts. Here, that TB-specific interferon-γ release assays might be useful for identifying individuals with progressive infections who are likely to develop the disease. This might provide an unprecedented advantage for TB control, namely targeted preventive therapy for individuals who are most at risk of developing active contagious TB.

5. CONCLUSION

Our framework can mine individual shift-invariant temporal patterns from heterogeneous events over different patient groups. The experimental results on both synthetic and real world electronic patient data are presented to demonstrate the effectiveness of the proposed method. The factors like location , age and gender decides the possibility of disease for a patient . Accuracy of data will be done up to nanoseconds which is really very important for the patients .Based on the clinical record the prediction of cancer type can be more accurate since it is based on the comparison of test report field values . And also temporal pattern generation enables the predictability of the stage of cancer growth . Though it is a challenging issue , it becomes possible for prediction using data mining techniques.

REFERENCES

[1] B. Cao, D. Shen, J.T. Sun, X. Wang, Q. Yang, and Z. Chen, "Detect and Track Latent Factors with Online Nonnegative Matrix Factorization," Proc. 20th Int'l Joint Conf. Artificial Intelligence, pp. 2689-2694, 2007.

[2] F.R.K. Chung, Spectral Graph Theory. Am. Math. Soc., 1997.

[3] C. Ding, T. Li, and M.I. Jordan, "Convex and Semi-Nonnegative Matrix Factorizations," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 32, no. 1, pp. 45-55, Jan. 2010.

[4] M. Dong, "A Tutorial on Nonlinear Time-Series Data Mining in Engineering Asset Health and Reliability Prediction: Concepts, Models, and Algorithms," Math. Problems in Eng., vol. 2010, pp. 1-23, 2010.

[5] J. Eggert and E. Korner, "Sparse Coding and NMF," Proc. IEEE Int'l Joint Conf. Neural Networks, vol. 2, pp. 2529-2533, 2004.

[6] W. Fei, L. Ping, and K. Christian, "Online Nonnegative Matrix Factorization for Document Clustering," Proc. 11th SIAM Int'l Conf. Data Mining, 2011.

[7] C. Fe´votte and J. Idier, Algorithms for Nonnegative Matrix Factorization with the Beta-Divergence, arXiv:1010.1763, 2010.

[8] P.O. Hoyer, "Non-Negative Matrix Factorization with Sparseness Constraints," J. Machine Learning Research, vol. 5, pp. 1457-1469, 2004.

[9] P.O. Hoyer, "Non-Negative Sparse Coding," Proc. 12th IEEE Workshop Neural Networks for Signal Processing, 2002.

[10] Y.R. Ramesh Kumar and P.A. Govardhan, "Stock Market Predictions—Integrating User Perception for Extracting Better Prediction a Framework," Int'l J. Eng. Science, vol. 2, no. 7, pp. 3305-3310, 2010.

[11] D.D. Lee and H.S. Seung, "Learning the Parts of Objects by Non- Negative Matrix Factorization," Nature, vol. 401, no. 6755, pp. 788-91, 1999.

[12] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms," Proc. Eighth ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery, pp. 2-11, 2003.

[13] J. Mairal, F. Bach Inria Willow Project-Team, and G. Sapiro, "Online Learning for Matrix Factorization and Sparse Coding," J. Machine Learning Research, vol. 11, pp. 19-60, 2010.

[14] F. Moerchen, "Time Series Knowledge Mining Fabian," PhD thesis, 2006.

[15] F. Moerchen and D. Fradkin, "Robust Mining of Time Intervals with Semi-Interval Partial Order Patterns," Proc. SIAM Conf. Data Mining, pp. 315-326, 2010.