

PAPER • OPEN ACCESS

Tree-like Dimensionality Reduction for Cancer-informatics

To cite this article: Xia Zhang *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **490** 042028

View the [article online](#) for updates and enhancements.



IOP | ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

Tree-like Dimensionality Reduction for Cancer-informatics

Xia Zhang ¹, Di Chang ², Weimin Qi ¹ and Zhiming Zhan ^{1,*}

¹School of Physics and Information Engineering, Jiangnan University, Wuhan, China, 430056

²Department of Computer Science, University of Georgia, Athens, Georgia, U.S., 30602

*Corresponding author e-mail: jasonzzm@tom.com

Abstract. Dimensionality reduction in machine learning currently has become a very heated research field. Traditional dimensionality reduction can be separated into two sub-fields of feature selection and feature extraction, but both of them are under local consideration. In this paper, an algorithm based on information theory, mutual information and maximum spanning tree will be proposed, in order to implement dimensionality reduction under global consideration rather than local consideration. Experimental results show it has a good performance, when the proposed algorithm is applied on gene sequences about cancer bioinformatics.

1. Introduction

1.1. Traditional Dimensionality Reduction

Nowadays machine learning has been a heated research area [1], where dimensionality reduction is a procedure to decrease the number of random variables by taking mathematic and statistic consideration [2] in order to conduct a small set of uncorrelated-level principle variables [3], which could mainly be separated into two subareas of feature selection and feature extraction [4].

Feature selection methods aim to extract a subset of the original input random variables, which of the term of “variable” can also be named as features or attributes. There are three most common approaches - wrapper approaches, filter, and embedding. Random variables are selected to be added in or removed out during machine modelling according the prediction error rate [5]. Especially in such cases, some traditional data analysis methods (e.g. regression or classification) can be done in advance in order to reduce hyperspace more accurately and efficiently [6] rather than in the original space.

Feature extraction is supposed to transform the data in the given original high-dimensional space to another space with lower dimensional density, in which data transformation can be linear such as in principal component analysis (PCA) [7], whereas a plenty of other non-linear dimensionality reduction approaches exist [8, 9] as well. Tensor representation was used to reduce dimensionality given by multidimensional data through multi-linear subspace machine learning [10].

One of the most common linear approaches for dimensionality reduction is to do maximization, which is popularly applied in PCA. It outputs a linear mapping function from the original data hyperspace into a space with fewer dimensions, in such a way which remains the diversity of the data in the low-dimensional representation as well. Practically, the covariance matrix of the data was constructed before the eigenvectors are computed based on the covariance matrix. The largest eigenvalues corresponding to the original data will be used in the reconstruction of a large segmentation of the diversity and variance. In addition, the first of some eigenvectors could be interpreted regard of the large-scale physical behavior of system, in which the given hyperspace as original input has already been decreased to a sub-space spanned by much fewer eigenvectors. PCA can be exploited technically in non-linear by the means of the “kernel trick”, whose experimental results proved its capability of constructing a non-linear mapping that maximizes variance and diversity. Current techniques have been proved to successfully learn the kernel via using semi-definite



programming rather than defining a fixed kernel, where maximum variance unfolding (MVU) is one of the most important approaches. The key point of MVU is to exactly preserve every pairwise distance between every nearest neighbor, and in the meanwhile, to maximize the distances between those points not being nearest neighbors. Neuroscience whose original datasets are with maximally informative dimensions goals to find out a lower-dimensional representation that preserves as much information as possible when compared to the original given data as input, in which of area MVU is sometimes used.

Another non-linear dimensionality reduction-method is, via involving autoencoders in, a special type of neural networks – named as “feed-forward” - which has a bottle-neck hidden layer [11]. To train such a deep encoders, a greedy-algorithm based layer-wise pre-training is typically exploited, which is then followed by a fine-tuning stage according to back-propagation.

As a high-dimensional dataset inputs, dimension reduction is often applied before using a K-nearest neighbor’s algorithm (k-NN) then for the purpose of minimize the side effects of dimensionality [12].

Feature extraction and dimension reduction usually can be combined within a step through using PCA, linear discriminant analysis (LDA), or canonical correlation analysis (CCA) as preprocessing, which is then followed by K-NN clustering on feature vectors within the hyperspace whose dimensionality has been already reduced, which of such process is called low-dimensional embedding [13] as well.

The only feasible options, in the context of very-high dimensional datasets, may be such as an approximate K-NN algorithm that searches "sketches" [14], random projection [15], local sensitive hashing, or other high-dimensional similar searching approaches according to the VLDB toolbox.

However, these traditional dimensionality reductions mentioned above are under “local” consideration rather than global consideration, which may bring to a local optimal solution instead of global.

1.2. *Cancer Bioinformatics*

Bioinformatics, especially for cancer informatics, is mainly based on the developments of mathematic, statistic, and computational approaches for the aim to process as well as analyze a large scale amount of biologic data. Particularly, cancer bioinformatics currently has been a crossover research area that intersects health care, computer science, together with information science, which is to acquire, store and use cancer data most thoroughly, efficiently, and effectively [16]. Technically, tools for cancer informatics research generally include clinical guidelines, computers, and information systems. The genomic revolution is impossible if we have not used sophisticated statistical algorithms on which micro-array expression profiling, DNA sequencing, or genomic sequence analysis. How to effectively integrate cancer informatics into realistic biology and how to train a new group of researchers especially for cancer- or bio-informaticists will be very challenged since a leadership with the crossover capability of biology plus information science must be mandatory.

Applications of bioinformatics will be crucial for the cancer treatment in the near future. A majority of cancer therapies work only for a small subset of cancer patients. It would be likely to retain true for a lot of molecularly-targeted drugs [17], which might cause a large percentage of cancer patients who are going to receive ineffective cancer treatments and which financially results in a huge problem in the health care system. To accurately develop tools for the purpose of delivering the right cancer treatment to the exact patient will be essential according to the biological features from each patient’s tumor. Being properly focused and supported, scientists have well developed novel gene-expression-based diagnostic tests nowadays in order to effectively assist cancer patients with huge difficulties of using currently existed cancer treatment approaches [18]. The continuous improvement of such tests will be an important segment of a novel paradigm for the future cancer therapeutics. In addition, cancer informatics will be crucial for discovering new drugs – especially for novel molecularly-targeted drugs with negative effects as small as possible for those cancer patients. A number of tumors have been proved consisting of mixtures of sub-clones which involve in different sets of mutated, overexpressed and silenced genes, which causes a huge difficulty of how to bio-technically identify a good molecular target.

Nowadays more studies focused translational research is important for all steps of such process [19], as multidisciplinary teams conducted the development and application of bioinformatics. In its broadest sense of taking advantage from bioinformatics, genomic technologies can be used to more effectively develop drugs as well as to more accurately target them to the right patients. These novel tools have achieved rapid advances in the context of cancer therapeutics in practical, which of therapy methods have been available and widely applied to many patients by now. The great progress is based on the proposal of innovative multi-disciplinary approaches which technically organize our wisdom for bringing a new generation of efficient and effective treatments for cancer disease.

2. Difficulties

The most important task for effective learning of Markov networks is to optimize joint probability distribution functions over random variables from observed data. Because such an optimization problem is computationally intractable, algorithms for Markov (Bayesian as well) network learning have resorted to heuristics or approximation methods [20]. Often such learning algorithms also assume constrained, typically tree or tree-like, topologies on the relationships of the random variables. While real-world networks are indeed mostly tree-like, such algorithms are often beyond theoretical rigorousness and without guarantees for the learning accuracy even upon the simplified tree-like topologies. As the solutions are established on non-global optimization, the learning performance may further deteriorate on big data that usually implicate large number of random variables and high dimensionality of data features.

2.1. Current Challenges

The most relevant notion for constraining tree-like network topologies is bounded tree width, evident by the fact that the almost all Markov network observed have small tree widths. The seminal work gave a rigorous argument that the tree topology constraint is a well-defined approximation for maximum likelihood Markov network learning and the best approximated network can be achieved by computing a maximum spanning tree. However, the situation becomes difficult when the constraint is relaxed to tree width $k \geq 1$. In particular, it is NP-hard to learn optimal Markov networks of tree width $k \geq 2$, even just for $k = 2$. The tractability issue of network learning with bounded tree width has attracted research with various proposed heuristic solutions, all but with guaranteed learning accuracy. Moreover, very limited work has been done on provable approximation algorithms, with one result of an $n^{O(k)}$ -time approximation algorithm. Nevertheless, the approximation ratio $(k + 1)!$ appears too large to be of practical usage.

Other than the obvious challenges in developing efficient accuracy-guaranteed learning algorithms, the research has yet to address another outstanding issue of how to translate correlations in a learned Markov network into causal relationships, even with tree width constrained topology. In particular, due to subtle differences between correlation and causality, it remains not clear how Markov network learning can be directly used to construct desired Bayesian networks. Due to the asymmetry nature of mutual information, the learned Markov networks are directed graphs, thus actually Bayesian trees.

2.2. Application on Cancer Bioinformatics

Datasets are supported by the CSBL, UGA, U.S. Data is generally described as that there are n genes and c conditions –modelling as n random variable and c dimensions, respectively. Each gene has m samples, which of every sample gives an expression level under each condition, where expression levels can be discretized to v values. In particular, m samples rather than a single one can give information about distribution and probability, which aims to reduce redundant conditions or dimensionalities among all of gene pairs.

3. Proposed Algorithm

We now consider a class of topological constraints as enforcement on approximation for the joint probability P . Let $X = \{X_1, X_2, \dots, X_n\}$ be a set of discrete random variables. The topology of X is a directed graph $G_X = (V, E)$ over vertices $V = \{1, 2, \dots, n\}$ such that for every pair $(i, j) \in V$, either (i, j) or (j, i) , but not both, belong to E . G_X is also called tournament.

Thus, we can measure difference between two distributions $P(X)$ and $P_G(X)$ with D_{KL} , Kullback-Leibler divergence. Minimizing D_{KL} will result in the maximum spanning tree problem, when G is assumed to be of tree topology [11]. In particular, if non-tree topology is desired, the problem will become computationally intractable, and relying on heuristics.

Assume we have a Markov tree T for variable $X = \{X_1, X_2, \dots, X_n\}$ with a root X_1 , where the tree topology is completely determined by π , the parent information. $P_T(X)$ will be conducted from (1).

$$P_T(X) = P(X_1) \prod_{i=2}^n P(X_i | X_{\pi(i)}) \quad (1)$$

Minimizing $D_{KL}(P(X), P_T(X))$ would tell us what T should be. Kullback-Leibler divergence is defined as: $D_{KL}(P(X), P_T(X)) = \sum_x P(x) \log_2 P(x) - \sum_x P(x) \log_2 P_T(x)$, where, $x = (x_1, \dots, x_n)$ is the vector of values for variables X_1, \dots, X_n .

In particular, take (1) into D_{KL} : $D_{KL}(P(X), P_T(X)) = -H(X) - \sum_x P(x) \log_2 P(X_1) \prod_{i=2}^n P(X_i | X_{\pi(i)})$

Consequently, Kullback-Leibler Divergence is (2).

$$D_{KL}(P(X), P_T(X)) = -H(X) + \sum_{i=1}^n H(X_i) - \sum_{i=2}^n I(X_i, X_{\pi(i)}) \quad (2)$$

Where, $I(X_i, X_{\pi(i)})$ refers to the Mutual Information between X_i and $X_{\pi(i)}$ (3).

$$I(X_i, X_{\pi(i)}) = \sum_{x_i, x_{\pi(i)}} p(x_i) \log \frac{P(x_i, x_{\pi(i)})}{p(x_i)p(x_{\pi(i)})} \quad (3)$$

Thus, because $-H(X) + \sum_{i=1}^n H(X_i)$ is always fixed according to the data, to minimize Kullback-

Leibler Divergence is to maximize $\sum_{i=2}^n I(X_i, X_{\pi(i)})$.

Hence, the algorithm can be described as:

- a) Construct graph G_X of n vertices, one for each variable $X_i \in X$;
- b) For every pair of (i, j) , edge (i, j) in G_X has a weight of $I(X_i, X_j)$;
- c) Find a maximum spanning tree T of G_X , where maximum spanning tree has the same algorithm as min spanning tree.

4. Experimental Results

Two gene data sets of cancer bioinformatics – downloaded from the sources mentioned in section 2.2 – which respectively are 487 genes from 6 pathways, and 523 genes from 17 more detailed pathways. After filtering, normalizing, discretizing, the first data set will remain 469 genes, and the second data set will remain 507 genes. Both two data sets are described in gene expression level of 12 iron-sulfur clustering metabolism, 19 oxidative stress, 65 protein damage response genes, 20 cell proliferation marker genes, and up to 500 cancer associated genes, whose mutual information shows in Figure 1&2.

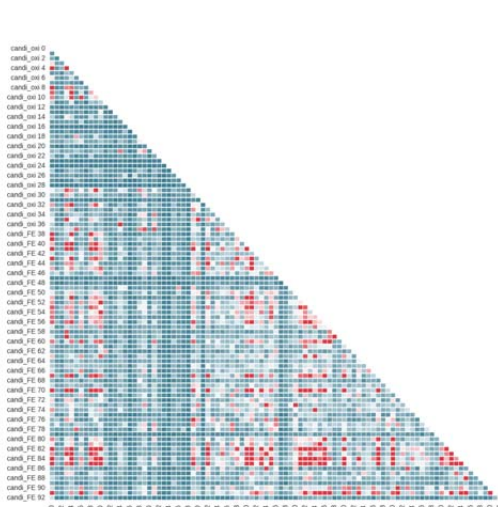


Figure 1. Pairwise Mutual Information Heatmap-diagram among Genes for Dataset #1

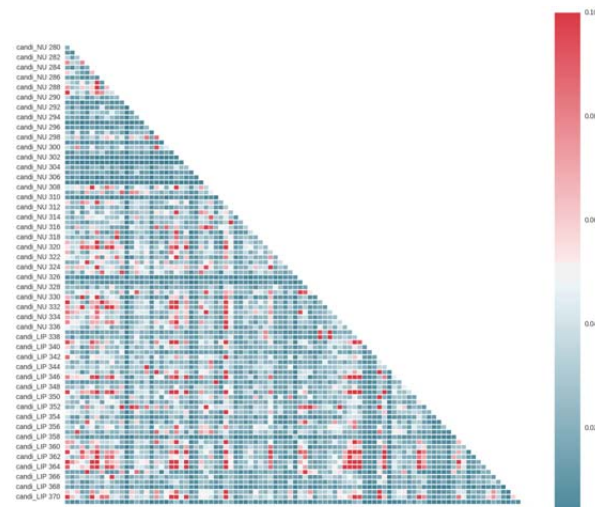


Figure 2. Pairwise Mutual Information Heatmap-diagram among Genes for Dataset #2

After processing the data set 1 with the proposed algorithm, a spanning tree is generated as Figure 3, where different colors may have their own clusters which are allocated in anywhere in the tree.

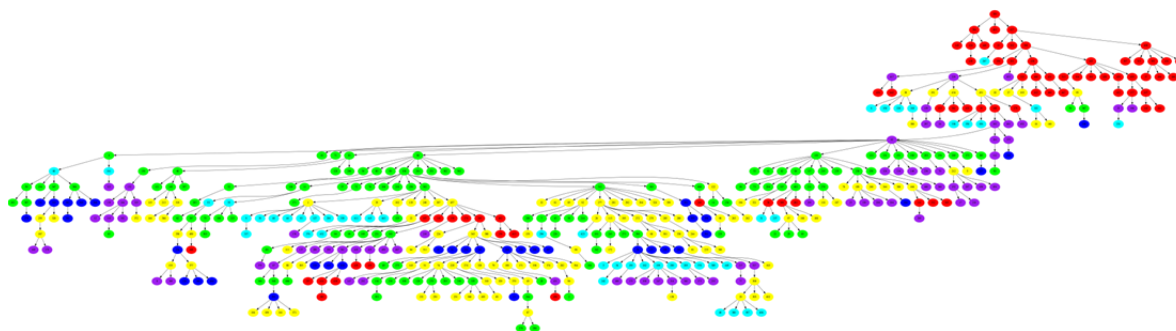


Figure 3. MI-based Spanning Tree with Clusters for Dataset 1

Figure 4 shows the remaining mutual information after a spanning tree with clustering is established.

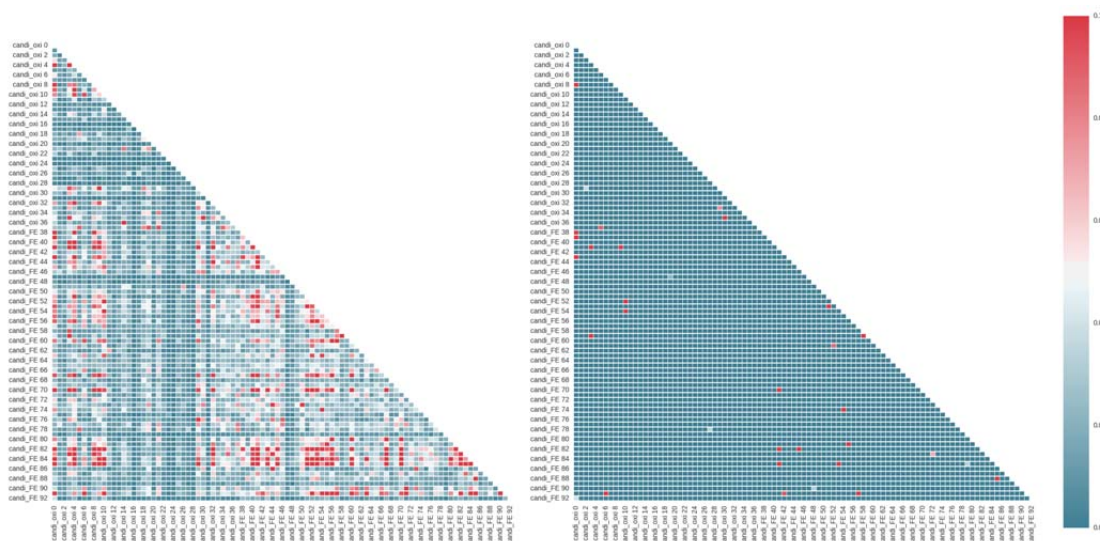


Figure 4. Heatmap-diagram of Remaining Mutual Information among Genes for Dataset #1

The total number of edges and nodes in the spanning tree are 468 and 469 respectively, where number of edges across different pathways, referring to higher correlation, is 311 (66.5% out of all edges). The result proves a good performance of only a few red points referring to correlation with high values of mutual information, although it could be further reduced if via k-tree instead of current spanning tree.

5. Conclusion

Traditional dimensionality reduction methods - whatever feature selection or feature extraction - are under local consideration. This paper proposed a tree-based algorithm by using information theory, mutual information and max spanning tree. Unlike traditional methods, the proposed algorithm will handle dimensionality reduction problem under global consideration instead of local consideration. Experiments show, the proposed algorithm performs well for the application of cancer bioinformatics.

References

- [1] D Chang, X Zhang, Q Liu, G Gao, Y Wu. Location based robust audio watermarking algorithm for social TV system. In Pacific-Rim Conference on Multimedia 2012 Dec 4 (pp. 726-738)
- [2] Roweis, S. T.; Saul, L. K. (2000). "Nonlinear Dimensionality Reduction by Locally Linear Embedding". *Science* 290 (5500): 2323–2326. doi:10.1126/science.290.5500.2323.
- [3] D Chang, X Zhang, Y Wu. A Multi-Source Steganography for Stereo Audio. *Journal of Wuhan University (Natural Science Edition)*. 2013;3: 277-284.
- [4] Pudil, P.; Novovičová, J. (1998). "Novel Methods for Feature Subset Selection with Respect to Problem Knowledge". In Liu, Huan; Motoda, Hiroshi. *Feature Extraction, Construction and Selection*. p. 101. doi:10.1007/978-1-4615-5725-8_7. ISBN 978-1-4613-7622-4.
- [5] Zhang X, Chang D. Sonic audio watermarking algorithm for cable-transmission. *The 2nd International Conference on Information Science and Engineering*, Vol. 7, 2010, pp. 5395-5398.
- [6] Zhang X, Chang D, Huang Q. An audio digital watermarking algorithm in DCT domain for air-channel transmitting. *Journal of University of Science and Technology of China*, 2001(41): 642-650.
- [7] Samet, H. *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann, 2006.
- [8] C. Ding, , X. He , H. Zha , H.D. Simon, *Adaptive Dimension Reduction for Clustering High Dimensional Data*, *Proceedings of International Conference on Data Mining*, 2002
- [9] Zhang X, Chang D, Guo W, etc. An Audio Steganography Algorithm Based on Air-Channel Transmitting. *Journal of Wuhan University (Natural Science Edition)*, 2011, 57(6): 499 – 505.
- [10] Lu, Haiping; Plataniotis, K.N.; Venetsanopoulos, A.N. (2011). "A Survey of Multilinear Subspace Learning for Tensor Data" (PDF). *Pattern Recognition* **44** (7): 1540 – 1551.
- [11] Hongbing Hu, Stephen A. Zahorian, (2010) "Dimensionality Reduction Methods for HMM Phonetic Recognition," *ICASSP 2010*, Dallas, TX
- [12] Kevin Beyer , Jonathan Goldstein , Raghu Ramakrishnan , Uri Shaft (1999) "When is “nearest neighbor” meaningful?". *Database Theory—ICDT99*, 217-235
- [13] Shaw, B.; Jebara, T. (2009). "Structure preserving embedding". *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09* (PDF): pp. 1. ISBN 9781605585161.
- [14] Shasha, D High (2004). *Performance Discovery in Time Series*. Berlin: Springer.
- [15] Bingham, E.; Mannila, H. (2001). "Random projection in dimensionality reduction". *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01*. p. 245. doi:10.1145/502512.502546. ISBN 158113391X.
- [16] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory*, 14:462 - 467, 1968.
- [17] G. Elidan and S. Gould. Learning Bounded Treewidth Bayesian Networks. *Journal of Machine Learning Research*, 9(2699-2731):2699-2731, 2008.
- [18] Xu, Y., J. Cui, and D. Puett, *Cancer bioinformatics*, Springer, 2014.
- [19] Tomczak, K., P. Czerwinska, and M. Wiznerowicz, *The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn)*, 2015. 19(1A): p. A68-77.
- [20] Zhang, X., et al. An audio digital watermarking algorithm transmitted via air channel in double DCT domain. *IEEE International Conference on Multimedia Technology*, 2011: pp. 2926-2930.