JIMMA UNIVERSITY

JIMMA INSTITUTE OF TECHNOLOGY

FACULTY OF COMPUTING

Program Computer Science

Data mining Assignment

Title: Random Matrix Theory

GROUP MEMBERS

NAME	IDNO
1. IYASU DEBELO……	RU0529/08
2. KELBESA MERGA……	RU4377/07
3. MELKAMU BEKELE…….	RU4401/07

Random matrix theory in data mining

Introduction

Random matrix theory is an important mathematical tool for statistical analysis of complex systems. Through the spectrum and Eigen state complex system of statistical analysis, random matrix theory obtained the random degree of actual data, which reveals the behavior characteristics of the actual data associated with the overall.

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facts and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.

Often the variables are not numbers but are instead qualitative descriptors called categorical data. We define and study similarity matrix, as a measure of similarity, for the case of categorical data. This is of interest due to a deluge of categorical data, such as movie ratings, top-10 rankings, and data from social media, in the public domain that require analysis.

We show that the statistical properties of the spectra of similarity matrices, constructed from categorical data, follow random matrix predictions with the dominant eigenvalue being an exception.

Many data mining applications deal with privacy sensitive data. Financial transactions, health-care records, and network communication traffic are some examples. Data mining in such privacy-sensitive domains is facing growing concerns. Therefore, we need to develop data mining techniques that are sensitive to the privacy issue.

This has fostered the development of a class of data mining algorithms that try to extract the data patterns without directly accessing the original data and guarantees that the mining process does not get sufficient information to reconstruct the original data. We consider class of techniques for privacy preserving data mining by randomly perturbing the data while preserving the underlying probabilistic properties.

It explores the random value perturbation-based approach a well-known technique for masking the data using random noise. This approach tries to preserve data privacy by adding random noise, while making sure that the random noise still preserves the "signal" from the data so that the patterns can still be accurately estimated.

It shows that in many cases, the original data (sometimes called "signal") can be accurately estimated from the perturbed data using a spectral filter that exploits some theoretical properties of random matrices. It presents the theoretical foundation and provides experimental results to support this claim.

Descriptions of Basic algorithms

Data mining is the process of extracting hidden patterns from data. But to analyze and handle all of this incoming data through data mining, we need to be able to separate the important information from the surrounding noise using different methods. This requires the use of increasingly sophisticated techniques.

We Investigates methods to make sense of huge data sets, to find the hidden correlations between apparently random pieces of information, their typical behavior, and random fluctuations. We consider things called matrices, where we have an array of data. So we take some data at random, put it in a big array of data, and then try to understand how to analyze it, for example to subtract the noise.

In probability theory and mathematical physics, a random matrix is a matrix-valued random variable. That is, a matrix in which some or all elements are random variables. Many important properties of physical systems can be represented mathematically as matrix problems. The field of random matrix theory in data analysis, has grown rapidly as the huge rise in the amount of data we produce. The theory is now used in statistics, finance, and telecommunications etc.

A common goal in data mining is to find regularities (or patterns) in the data a random matrix is a matrix of given type and size whose entries consist of random numbers from some specified distribution. In data analysis we motivate to use random matrix for the following reasons:

- ✓ understanding large deviations

- ✓ The probability of finding unlikely events or unusual behavior within the array of data and in connecting the theory with that of topological expansion, in which random matrices are used to help solve combinatorial questions.
- ✓ A dataset contains a certain amount of information
- ✓ A random dataset has high entropy
- ✓ Work towards reducing the amount of entropy in the data
- ✓ Alternatively, increase the amount of information exhibited by the data
- ✓ In deep learning neural network configurations with random weights play an important role in the analysis of dataset.
- ✓ They define the initial loss landscape and are closely related to kernel and random feature methods.

One focus area in mathematical physics involves studying the spacing of ordered entities in a system. Suppose there is a system with the observables {e1, e2, e3…., en}, where e1 ≤ e2 ≤ e3 ≤ . . . ≤ en, which describe some property of that system. The spacing between the distinct ei could be a central question to explore.

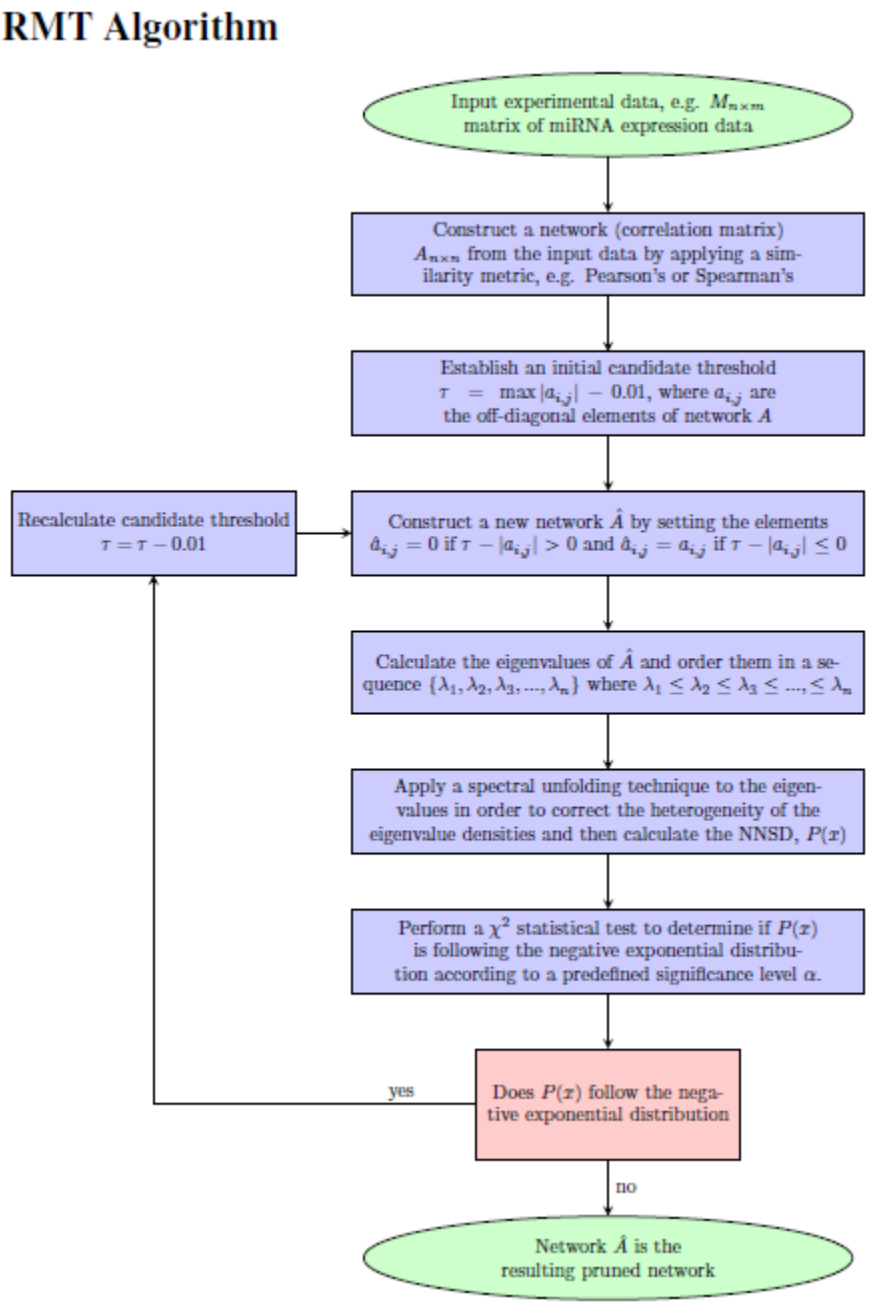
Which includes Examining the spacing between different energy levels of complex nuclei prime gaps, and the ordered arrangement of the eigenvalues of symmetric matrices. Having a full, comprehensive body of knowledge about the system itself would lend 18 to an exact knowledge of the spacing between each ei and ei+1.

When studying complex systems whose interrelationships quickly become too thorny to grasp, the converse approach is often taken; namely ei to gain more of an understanding about the underlying system from the knowledge of how they are spaced.

RMT is an approach for studying massive, complex systems that are too difficult to directly analyze because of their labyrinth of connected components. The main functionality of RMT is to contribute to an understanding of complex systems primarily by analyzing the statistical properties of eigenvalue spacing distributions of the systems.

By examining the distribution of the eigenvalue spacing of a network and comparing that distribution to known global properties associated with some universal system, one can detect system specific behavior and determine what is unique about the network that is being analyzed.

The behavior just described is called universality, a key concept to the mechanics of RMT. Universality is the observation that there are behaviors exhibited in a large collection of systems that share some similar characteristics, which are independent of the individual systems.



RMT is a thresholding technique that is knowledge-independent. This means that the algorithm is not affected by what the actual data is measuring. The threshold is a cutoff number between two units or elements that determines if they are correlated or not. Establishes an initial threshold whose value is significantly larger than most of the other values in the matrix. This step of establishing the first threshold value is somewhat arbitrary. However, the value needs to be high enough so that one can be fairly certain that anything with a value above this threshold (in absolute value) will actually be correlated and not just have coincidental correlation.

Large random matrices are used models for the massive data arising from the monitoring of the massive input and massive output system. Data Modeling with Large Random Matrices Naturally, we assume n observations of p-dimensional random vectors $x1,x2,...,xn \in Cp\times1$. We form the data matrix $X = (x1,x2...,xn) \in Cp\times n$, which naturally, is a random matrix due to the presence of ubiquitous noise.

While n is assumed to be arbitrary. The possibility of arbitrary sample size n makes the classical statistical tools infeasible. We are asked to consider the asymptotic regime $p \to \infty, n \to \infty, p/n \to c \in (0,\infty)$, (1) while the classical regime considers p fixed, $n \to \infty, p/n \to 0$

Matrices in data mining

	Bread	Butter	Beer
Anna	1	1	0
Bob	1	1	1
Charlie	0	1	1

Customer transactions

	Data	Matrix	Mining
Book 1	5	0	3
Book 2	0	0	7
Book 3	4	6	5

Document-term matrix

	Avatar	The Matrix	Up
Alice		4	2
Bob	3	2	
Charlie	5		3

Incomplete rating matrix

	Jan	Jun	Sep
Saarbrücken	1	11	10
Helsinki	6.5	10.9	8.7
Cape Town	15.7	7.8	8.7

Cities and monthly temperatures

Wigner Matrices. The most basic model for a random matrix ensemble is the Wigner matrix ensemble. For the sake of clarity, an ensemble of random matrices is a family, group, or collection of random matrices where any member of the infinite group of matrices can represent a state of the entire group. Wigner matrices are historically important because they were the first model of a random matrix ensemble.

Now to define a Wigner matrix (ensemble): specifically a real Wigner matrix which is a Wigner matrix with real-number entries. Without loss of generality, when referring to a real Wigner matrix it will be referred to simply as a Wigner matrix

Suggestion

Randomized algorithms for very large matrix problems have received a great deal of attention in recent years. Much of this work was motivated by problems in large-scale data analysis, largely since matrices are popular structures with which to model data drawn from a wide range of application domains, and this work was performed by individuals from many different research communities. While the most obvious benefit of randomization is that it can lead to faster algorithms, either in worst-case asymptotic theory and/or numerical implementation, there are numerous other benefits that are at least as important.

For example, the use of randomization can lead to simpler algorithms that are easier to analyze or reason about when applied in counterintuitive settings; it can lead to algorithms with more interpretable output, which is of interest in applications where analyst time rather than just computational time is of interest; it can lead implicitly to regularization and more robust output; and randomized algorithms can often be organized to exploit modern computational architectures better than classical numerical methods.

This monograph will provide a detailed overview of recent work on the theory of randomized matrix algorithms as well as the application of those ideas to the solution of practical problems in large-scale data analysis. Throughout this review, an emphasis will be placed on a few simple core ideas that underlie not only recent theoretical advances but also the usefulness of these tools in large-scale data applications.

Crucial in this context is the connection with the concept of statistical leverage. This concept has long been used in statistical regression diagnostics to identify outliers; and it has recently proved crucial in the development of improved worst-case matrix algorithms that are also amenable to high-quality numerical implementation and that are useful to domain scientists. This connection arises naturally when one explicitly decouples the effect of randomization in these matrix algorithms from the underlying linear algebraic structure. This decoupling also permits much finer control in the application of randomization, as well as the easier exploitation of domain knowledge. Most of the review will focus on random sampling algorithms and random projection algorithms for versions of the linear least-squares problem and the low-rank matrix approximation problem.

These two problems are fundamental in theory and ubiquitous in practice. Randomized methods solve these problems by constructing and operating on a randomized sketch of the input matrix A | for random sampling methods, the sketch consists of a small number of carefully-sampled and rescaled columns/rows of A, while for random projection methods, the sketch consists of a small number of linear combinations of the columns/rows of A.
Depending on the specifics of the situation, when compared with the best previously-existing deterministic algorithms, the resulting randomized algorithms have worst-case running time that is asymptotically faster; their numerical implementations are faster in terms of clock-time; or they can be implemented in parallel computing environments where existing numerical algorithms fail to run at all.

Random matrix theory (RMT) is an area of study that has applications in a wide
Variety of scientific disciplines. The foundation of RMT is based on the analysis of the
Eigenvalue behavior of matrices. The eigenvalues of a random matrix (a matrix with
Stochastic entries) will behave differently than the eigenvalues from a matrix with nonrandom properties. Studying this two of the eigenvalue behavior provides the means to which system-specific signals can be distinguished from randomness.
In particular, RMT provides an algorithmic approach to objectively remove noise from matrices with embedded signals. The function of the RMT algorithm used for data filtering is described. A survey of network analysis tools is also included as a way to provide insight on how to begin a

rigorous, mathematical analysis of networks. Furthermore, the results of applying the RMT algorithm to a set of data is provided.

The results of applying the RMT algorithm to the data are provided along with an implementation of the resulting network into a network analysis tool. These preliminary results demonstrate the facility of RMT coupled with network analysis tools as a basis for data.

5. CONCLUSIONS

Random matrix theory (RMT) has been shown to be a reliable method for constructing meaningful networks in a variety of data. It has also been shown to be an effective tool for predicting functions such as biologicals of genes.

We have reported some initial demonstrations of the theoretical framework: the massive amount of data can be naturally represented by (large) random matrices. Although intuitive, the systematic use of this framework is relatively recent. It appears that this work may be the first attempt to investigate the empirical science, in order to quantify the accuracy of the theoretical predictions with experimental findings.

RMT was explored as a network pruning tool because it is a way to objectively find a threshold of correlation in a network. Network analysis fills a crucial need in biology; however, to analyze the networks, they must be filtered to some extent.

This data filtering, or network pruning, process is necessary in order to remove spurious correlations from the data. The choice of a threshold value in which to eliminate the noise in a network greatly influences the resulting network.

The choice of a threshold value impacts the sensitivity and specificity of the node connections. Therefore, an omnipresent question encountered by scientists is what the most appropriate threshold value to use is. If the threshold value chosen is too high, then the computational complexity is lowered at the expense of losing node connections.

The resulting network may therefore be too sparse and information could be lost. On the other hand, if the threshold value is too low, then information is not lost but the computational requirements are increased. Too many spurious correlations may appear which could mask the true signal from emerging. RMT provides a way to objectively choose the most appropriate threshold value.

The use of RMT was investigated by applying the RMT algorithm, incorporated in the RMTGeneNet software, to an open-source miRNA data set. Preliminary results and figures were provided to demonstrate the performance of RMT. The preliminary results are encouraging and leave room for much future work to be done.