Exploring Data Mining Research in Nanotechnology

Aparna S. Varde [1], Jianyu Liang [2], Elke A. Rundensteiner [1] and Richard D. Sisson Jr. [2]
1. Department of Computer Science
2. Department of Mechanical Engineering
Worcester Polytechnic Institute, Worcester, MA 01609
(aparna | jianyul | rundenst | sisson)@wpi.edu

Abstract

Scientific datasets often consist of complex data types such as images. Mining such data presents interesting issues related to semantics. In this paper, we explore the research issues in mining data from the field of nanotechnology. More specifically, we focus on a problem that relates to image comparison of material nanostructures. A significant challenge here relates to the notion of similarity between the images. Features such as size and height of nano-particles and inter-particle distance are important in image similarity as conveyed by domain experts. However, there are no precise notions of similarity defined apriori. Hence there is a need for learning similarity measures. In this paper, we describe our proposed approach to learn similarity measures for graphical data. We discuss this with reference to nanostructure images. Other challenges in image comparison are also outlined. The use of this research is discussed with respect to targeted applications.

1. Introduction

In recent years there has been much interest mining scientific datasets [3, 5, 8, 11, 15]. This presents several challenges pertaining to the complexity of the data types and the semantics of the domain. Scientific data often consists of images which have to be interpreted with reference to context. Discovering knowledge from such data presents issues related notions of similarity, interestingness measures and visualization of the data mining results.

In this paper, we explore such issues in the context of nanotechnology, a popular area in scientific databases today. The field of nanotechnology relates to the design, characterization, production and application of structures, devices and systems by controlling shape, size, structure and chemistry of materials at the nanoscale level. It deals with the understanding and control of matter at dimensions of roughly 1 to 100 nanometers, where unique phenomena enable novel applications [13]. Nanotechnology involves a confluence of several disciplines such as physics, chemistry, biology and materials science.

Data from nanotechnology, as in the case of any scientific domain, is of various types such as numbers, plain text, graphs and images. In this paper we focus on images depicting nanostructures of materials. An interesting problem is the comparison of such images in computational analysis. The inferences drawn from comparison are useful in real-world applications such as materials science, biomedicine and tissue engineering [2]. To enable effective comparison, it is essential to preserve the semantics of the images. Accordingly, it is important to define notions of similarity and interestingness measures for comparison with respect to the domain. Moreover, visualization of image comparison results taking into account user interests is also an issue.

In this paper, we focus on one issue, namely, the notion of similarity or distance between the images. Domain experts are able to identify some features crucial in image comparison. For example, the size of the nano-particles within the image, the distance between these nano-particles and the height of the nano-particles in the cross-section of the image are considered to be significant. However, the experts have only subjective notions of similarity, not a precise measure. Hence there is need to learn similarity measures for such images. We describe our proposed approach called LearnMet [14] that has been used in a computational estimation system [15] to learn distance metrics for graphical data. We discuss the issues in enhancing this approach for images. We also discuss some of the other challenges in image comparison.

The rest of this paper is organized as follows. Section 2 gives a background of the domain and the motivation for the given problem. Section 3 describes in detail the problem of comparing nanostructure images along with its associated challenges. Section 4 discusses one particular challenge related to the notion of similarity between the images and a potential method of addressing it. Section 5 summarizes related work. Section 6 states the conclusions.

2. Background and Motivation

The investigation of cell-substrate interactions plays an important role in biomedical and tissue engineering research efforts. Understanding how cells interact with substrates will lead to the ability to optimize substrates for specific biomedical applications [2]. Studies have shown that microscale topography influences cells to assume the shape of underlying patterns and form cytoskeletons oriented to the patterns [4]. There are few studies which have analyzed cell interactions with nanostructured substrates [17].

In order to extensively analyze cell-substrate interactions on the nanoscale, simple, inexpensive, and scalable nanofabrication methods which can accommodate a wide variety of materials must be developed. It is necessary to make this step from the microscale to the nanoscale. Based upon the need for more thorough nanoscale research there is development of simple, inexpensive, and scalable nanofabrication methods which accommodate different types of materials. Bone cells are being cultured on the nanostructures and the cell adhesion, proliferation, differentiation, and mineralization are being monitored using standard cell culture arrays and electrochemical impedance spectroscopy [2].

The results of this research will facilitate the fabrication of biological nanostructures and contribute to the continuing efforts to understand how cells function in the presence of synthetic substrates for biomedical and tissue engineering applications [2, 4].

An important step in this research is the comparison between images depicting the cell responds to various nanostructures used in the given applications. This comparison enables drawing inferences about the impact of the nanostructure on the cells. For example, image comparison at different stages of cell culturing is very important to understand the interaction of the cells with nanostructures. Comparison of different nanostructures at the same stage but obtained under different cell culturing conditions helps to determine how the inter-cellular interactions are affected with the existence of various nanostructures. .

The research has potential use in targeted applications. Some of these applications include investigating the adhesion between cells and substrates in biomedical data, studying the alignment of cells and the differentiation between cells. This caters to the broader goals of developing materials for implants in the human body and helping the human skin to heal.

It is desirable to automate the comparison between the nanostructure images for computational analysis. The comparison can be automated using techniques such as clustering [9] and similarity search [8]. However, in order to achieve effective comparison, it is essential to capture the semantics of the images. In other words, it is important to make the comparisons analogous to a domain expert.

The problem of image comparison facing the nanoscience community thus opens potential avenues for data mining research. This problem is discussed in detail in this paper.

3. Comparison of Nanostructure Images

3.1. Goals of Comparison

A nanostructure is a structure with arrangement of its parts in the nanometre scale. Nanostructures of materials are observed to study their properties [13]. Comparing nanostructure images enables us to determine whether the nanostructure play a crucial role in the cells. It helps to answer questions such as:

- What is the difference in nanostructure at various locations of a given sample?
- How does the nanostructure evolve at different stages of a physical / chemical / biochemical process?
- To what extent does processing under different conditions affect the cell-nanostructure interaction at the same stage of a process, such as cell culturing?

This is explained with reference to the figures below depicting images of nanostructures taken with a Scanning Electron Microscope (SEM).

Figure 1 shows a top view of a silicon nanopillar array [2]. Figure 2 is a top view of the same specimen at a different location and more zoomed in. Figure 3 is a top view of a nanopore array etched into silicon [2].

From these images it is observed that Figures 1 and 2 depict different nanostructures due to the difference in location and in the level of zooming. Figures 1 and 3 on the other hand show different nanostructures based on the conditions of the physical process used to obtain them.

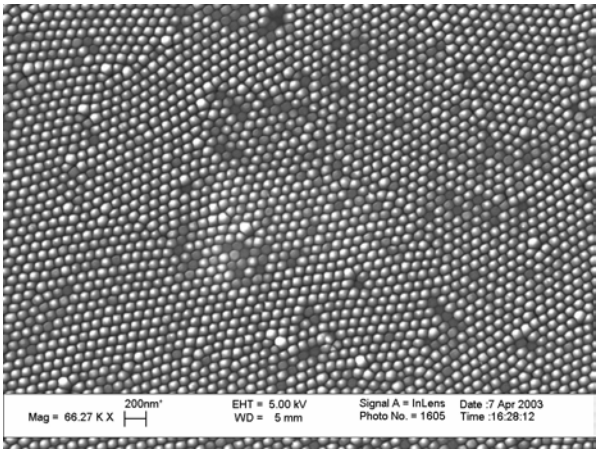

Figure 1: Top view of Si nanopillar array

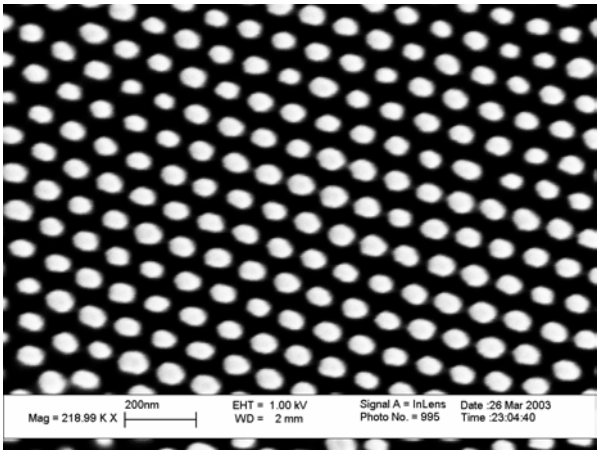

Figure 2: Top view of Si nanopillar array at a different location of the given sample and more zoomed in

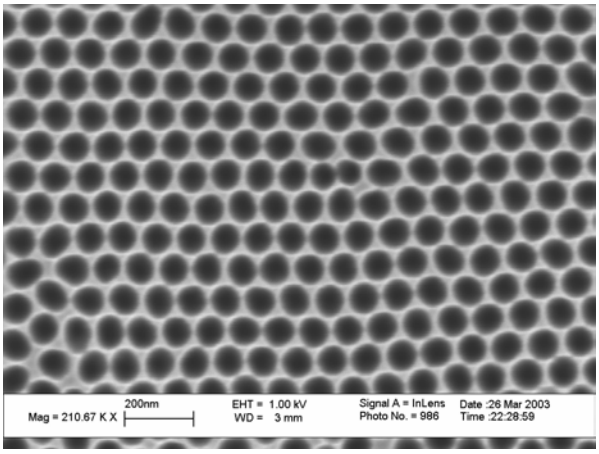

Figure 3: Top view of nanopore array etched into Si

3.2. Issues in Comparison

In order to make nanostructure comparisons, domain experts typically observe certain features of the images, such as:

- Nanoparticle size: This refers to the dimensions of each individual particle in the nanostructure.
- Inter-particle distance: This is the average distance between the particles as seen in a 2-dimensional space.
- Nanoparticle height: This indicates to what extent the particles project above the surface in a cross-section and is recorded as additional data.

When experts manually make such comparisons, these are the subjective notions of similarity. The greater the similarity between these individual features, the greater is the similarity between the nanostructure images as a whole. Thus if two images have the same nanoparticle size, the same inter-particle distance and the same nanoparticle height, then they would be considered similar during visual inspection by domain experts. Also, the experts would manually take into account the effect of aspects such as the level of zooming and the nature of the cross-section (top-view, oblique view etc.) in making such comparisons.

Thus in order to automate image comparison for computational analysis, it is useful to incorporate the reasoning of the experts. However, it is to be noted that this notion of similarity is subjective which is acceptable for visual inspection. In computational analysis, there is a need for objective similarity measures in order to compare these images using processes such as clustering [9]. This motivates the need for learning such domain-specific similarity measures for images.

Another important issue in image comparison is to define interestingness measures. Some knowledge discovered from the comparison can be obvious with respect to the domain. Other knowledge may be less obvious but may not provide any useful information. Thus, based on such criteria, it is essential to define what is interesting to targeted users. These measures again need to be objective so as to facilitate computational analysis. For example, in data mining techniques such as association rules, common interestingness measures are rule confidence and rule support [6]. Likewise, there is a need to define interestingness measures in image comparison.

Having performed analysis by data mining techniques such as similarity search [8] and clustering

[9], it is desirable to effectively visualize the data mining results. For instance, users may be interested in observing how a particular feature such as nanoparticle size varies from one specimen to another in evolutions of a physical process such as etching [2].

One possible way to address this would be to model each feature as an *attribute* the content of the feature as the *value* of that attribute. Thus, for example, "nanoparticle size" could be an attribute and "200 nanometers" could be its value.

Tools such as XMDV [16] incorporating techniques such as parallel co-ordinate plots and star-glyphs plots for visualizing multivariate data could then be used. Figure 4 shows an example of a star-glyphs plot for graphical data. Each vertex represents an attribute and the distance from the center of the star represents its value. The number of attributes and their combinations can be customized according to user preferences. Clusters and similarities can be visualized by comparing their shapes and sizes [16].

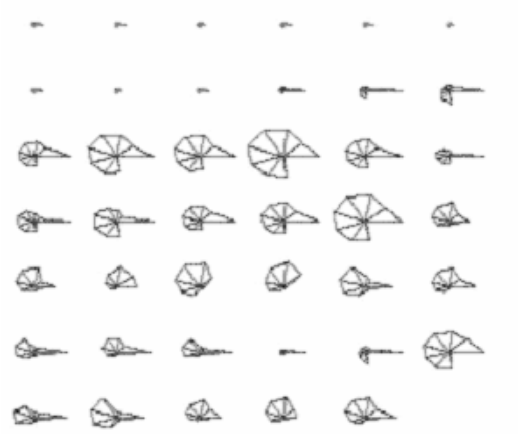

Figure 4: Example of Star Glyphs Plot

However, such visualization for images poses issues such as feature selection, data post-processing and adaptation of existing techniques to enable the visualization. Addressing these poses challenges.

Hence, in general the following issues can be outlined in comparison of nanostructure images.

- Learning a notion of similarity for the nanostructures
- Defining interestingness measures for comparison between nanostructures
- Visualizing the results of comparison based on user interests

We elaborate on one of these issues, namely, the notion of similarity between the nanostructure images.

4. Notion of Similarity

The problem of similarity measures for complex data types has been approached in several ways [1, 7, 8, 12, 14, 19].

Our earlier work, LearnMet [14] learns domain-specific distance metrics for graphical data. More specifically, we deal with 2-dimensional graphical plots of scientific functions. These graphs plot a dependent versus an independent variable depicting the behavior of process parameters. The graphs have semantics associated with them related to features such as the absolute position of points, statistical observations and critical regions. LearnMet learns distance metrics to capture the semantics of these graphs taking into account such features. We briefly describe the LearnMet approach to learn distance metrics for graphical data and discuss this in the context of images.

4.1. LearnMet Approach for Graphical Data

In the LearnMet approach experts provide actual clusters depicting the notion of correctness in the domain. A LearnMet distance metric is defined as a weighted sun of individual metrics such as position-based, statistical or critical distances [14].

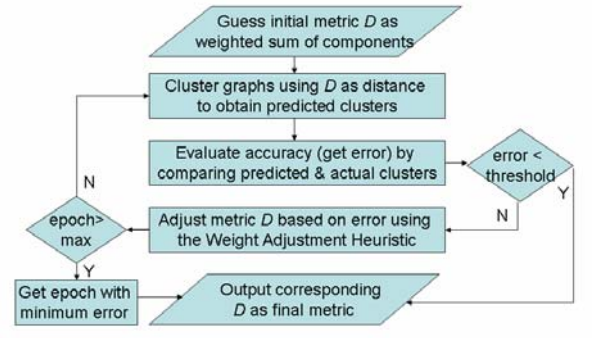

Figure 5: The LearnMet Approach

In the learning process, predicted clusters obtained with a guessed initial LearnMet metric are iteratively compared with the actual clusters. In each iteration, the metric is refined until the error between predicted and actual clusters is minimal or below threshold. The metric corresponding to the lowest error is the learned metric. This approach is illustrated in the flowchart in Figure 5.

The basic idea in LearnMet can be applied to learn similarity measures for images. However, there are

several issues involved here as discussed in the next subsection.

4.2. Learning Similarity Measures for Images

A 3-dimensional nanostructure image is more complex than a 2-dimensional graphical plot. The semantics associated with the images is of a different nature. Some of the features applicable to images can be explicitly identified such as nanoparticle size, interparticle distance and nanoparticle height. However, some aspects are more subtle such as the level of zooming involved in producing the images and the nature of the cross-section for the nanostructure sample. Even among the features identified, the relative importance is not defined apriori. Moreover, it is not always feasible for domain experts to provide actual clusters of images depicting the notion of correctness.

Given these issues, it does not seem feasible to directly apply the LearnMet approach for nanostructure images. Considerable enhancement is needed to learn a notion of similarity for images.

Thus, in general it is required that the learning of similarity measures for nanostructure images be done under the following conditions:

- Some features of the image are explicitly defined while others are more subtle
- Relative importance of the features is not known apriori
- Actual clusters of images are not provided by domain experts

We propose to address this learning using an approach that involves minimizing the intra-cluster distance and maximizing the inter-cluster distance for each cluster of images. The goal is to learn a similarity measure that achieves this.

The minimum description length principle [10] that minimizes the sum of encoding a theory and examples using the theory is likely to be useful here. The theory in our context could be a cluster representative while the examples could be all the other images in the cluster. This approach involves several challenges such as defining heuristics for iteratively adjusting the similarity measure until the intra-cluster distance is minimal.

Thus, the learning of similarity measures for nanostructure images is proposed to be done in an iterative manner. This forms a topic of our ongoing research.

5. Related Work

Several similarity measures exist in the literature such as Euclidean and Manhattan distances, [6] Tri-plot-based measures [12], Edit distances [1] and order-based measures [6, 8, 18]. However it is not known in advance which of these if any apply in the context of the given problem.

In [7] they propose a distance learning method to find the relative importance of dimensions for n-dimensional objects. However, their focus is on dimensionality reduction and not on semantics. In [19] they learn which type of position-based distance is applicable for the given data starting from the formula of Mahalanobis distance. However they do not deal with other distance types concerning images and semantics. In [8] they overview various distance types for similarity search over multimedia databases. However no single measure encompassing several types is proposed.

Interestingness measures have been defined in the literature for data mining techniques such as association analysis, clustering and classification [6, 9, 18]. However, there is often a need for interestingness measures to be domain-specific and need to cater to interests of various users. Hence we need to define such measures in the context of our problem.

The XMDV tool [16] incorporates techniques such as parallel co-ordinates plots and star-glyphs plots for visualization of multivariate data. However, the adaptation of this approach to the given problem involves several issues.

6. Conclusions

This paper describes the research issues in mining nanostructure images. The goal is to compare the nanostructures to analyze material properties. The issues in mining relate to the notion of similarity between the images, the interestingness measures in comparison and the visualization of the mining results. This research benefits the nanoscience community. The broader goal is to study properties of materials at the nanoscale level in order to solve problems in the fields of materials science, biomedicine and tissue engineering.

Acknowledgments

The authors thank the researchers from the Nanofacbrication and Nanomanufacturing group in the Department of Mechanical Engineering at WPI.

References

[1] L. Chen and R. Ng, "On the Marriage of Lp-Norm and Edit Distance", *VLDB*, Toronto, Canada, Aug 2004, pp. 792 - 803.

[2] S. A. Dougherty, J. Liang, G. D. Pins, *Preceision Nanostructure Fabrication for the Investigation of Cell Substrate Interactions*, Technical Report, Worcester Polytechnic Institute, Worcester, MA, June 2006.

[3] U. Fayyad, D. Haussler, and P. Storoltz. Mining Scientific Data, *Communications of the ACM*, Vol. 39, No. 11, pp. 51--57, Nov 1996.

[4] R.G. Flemming, C.J. Murphy, G.A. Abrams, S. L. Goodman and P.F. Nealey, "Effects of synthetic micro and nano-structured surfaces on cell behavior", *Biomaterials*, 1999, Vol. 20, pp. 573-588.

[5] L. Gao and X. S. Wang, "Continually Evaluating Similarity-Based Pattern Queries on a Streaming Time Series", *SIGMOD,* Madison, WI, 2002, pp. 370-381.

[6] J. Han and M Kamber, *Data Mining: Concepts and Techniques,* Morgan Kaufmann, CA, 2001.

[7] A. Hinneburg, C. Aggarwal and D. Keim, What is the Nearest Neighbor in High Dimensional Spaces, *VLDB*, 506 – 515. Cairo, Egypt. August 2000.

[8] D. Keim and B. Bustos, "Similarity Search in Multimedia Databases", *ICDE,* Boston, MA, Mar 2004, pp. 873 – 874.

[9] Kaufman and Rosseau. *Finding Groups in Data: An Introduction to Cluster Analysis,* John Wiley, NY, 1988.

[10] J. Rissanen, "Stochastic Complexity and the MDL Principle". *Econometric Reviews,* 1987, Vol. 6, pp. 85-102.

[11] Y. Reich and S. V. Barai, "Evaluating Machine Learning Models for Engineering Problems", *Artificial Intelligence in Engineering,* 1999, Vol, 13, pp. 257 – 272.

[12] A. Traina, C. Traina, S. Papadimitriou and C. Faloutsos, "TriPlots: Scalable Tools for Multidimensional Data Mining", *KDD*, San Francisco, CA, 2001, pp. 184 – 193.

[13] United States National Nanotechnology Initiative, Technical Report, Feb 2006.

[14] A. S. Varde, E. A. Rundensteiner, C. Ruiz, M. Maniruzzaman and R. D. Sisson Jr. "Learning Semantics-Preserving Distance Metrics for Clustering Graphical Data", *KDD's MDM*, Chicago, IL, Aug 2005, pp.107 – 112.

[15] A. S. Varde, E. A. Rundensteiner, C. Ruiz, D. C. Brown, M. Maniruzzaman and R. D. Sisson Jr. "Integrating Clustering and Classification for Estimating Process Variables in Materials Science" *AAAI Poster Track,* Boston, MA, July 2006.

[16] M. O. Ward, "XMDV Tool: Integrating Multiple Methods for Visualizing Multivariate Data", *Visualization,* Washington, DC, Oct 1994, pp. 326 – 333.

[17] T.J. Webster and T.A. Smith, "Increased Osteoblast Function on PLGA Composites Containing Nanophase Titania", *Journal of Biomedical Materials Research,* 2005, Vol. 74A, pp. 677-686.

[18] I. Witten and E. Frank *Data Mining: Practical Machine Learning Algorithms with Java Implementations*, Morgan Kaufmann Publishers, CA, USA, 2000.

[19] E. Xing, A. Ng, M. Jordan, S. Russell, "Distance Metric Learning with Application to Clustering with Side Information", *NIPS*, Vancouver, Canada, Dec 03, pp. 503-512.