The Use of Nonlinear Manifold Learning in Recommender Systems

Aleksandra Klašnja-Milićević,
High Business school, University of Novi Sad, Serbia
aklasnja@yahoo.com

Mirjana Ivanović
Faculty of Science, Department of Mathematics and Informatics, University of Novi Sad, Serbia
mira@dmi.uns.ac.rs

Alexandros Nanopoulos
Information Systems and Machine Learning Lab, University of Hildesheim, Germany
nanopoulos@ismll.de

ABSTRACT

Recommender systems have recently gained much attention as a new business intelligence tool for e-commerce. Applying a recommender system for an online retailer store helps to enhance the quality of service for customers and increase the sale of products and services. One successful recommender system technology is collaborative filtering which employs statistical techniques to find a set of users known as neighbors, who have a history of agreeing with the target user. Collaborative filtering has been shown to produce high quality recommendations, but the performance degrades for a large number of users and products. The problems associated with high dimensionality in the recommender systems have been discussed in several studies. It has been reported that dimensionality reduction techniques are effective for k-NN algorithms used typically for collaborative filtering. Non Linear Dimensionality Reduction (NLDR) techniques in turn have performed better than linear dimensionality reduction techniques. However, manifold alignment for the purposes of collaborative filtering couldn't be used as an effective approach. In this paper, some limitations of the use of nonlinear manifold learning for collaborative filtering will be considered.

Key Words: Recommender System, Non Linear Dimensionality Reduction (NLDR), Isomap, LLE

1. Introduction

Recommender systems apply knowledge discovery techniques to the problem of making product recommendations during a live customer interaction [1]. These systems are achieving widespread success in E-commerce nowadays, especially with the advantage of the Internet. One successful recommender system technology is collaborative filtering, which works by matching customer preferences to other customers in order to make appropriate recommendations. Collaborative filtering (CF) has been shown to produce high quality recommendations, but the performance degrades with the number of customers and products.

An E-commerce recommender system may easily involve millions of customers and products. This amount of data poses a great challenge to the CF algorithms in that the recommendations need to be generated in real-time. Furthermore, the algorithm also has to cope with a steady influx of new users and items. For the majority of the algorithms proposed to date, the primary emphasis has been given into improving recommendation

accuracy. While accuracy is certainly important and can affect the profitability of the company, the operator simply cannot deploy the system if it does not scale to the vast data of the site. The tremendous growth of customers and products poses two key challenges for recommender systems [2]. The first challenge is to improve the quality of the recommendations for the consumers. Consumers need recommendations they can trust to help them find products they will like. If a consumer trusts a recommender system, purchases a product, and finds out he does not like the product, the consumer will be unlikely to use the recommender system again. Another challenge is to improve the scalability of the collaborative filtering algorithms. These algorithms are able to search tens of thousands of potential neighbors in real-time, but the demands of modern E-commerce systems are to search tens of millions of potential neighbors. In some ways these two challenges are in conflict, since the less time an algorithm spends searching for neighbors, the more scalable it will be, and the worse its quality. For this reason, it is important to treat the two challenges simultaneously so the solutions discovered are both useful and practical.

New technologies are needed for dramatically improving the scalability of recommender systems. It has been reported that dimensionality reduction techniques are effective for k-NN algorithms used typically for collaborative filtering [3]. Non Linear Dimensionality Reduction (NLDR) techniques in turn have performed better than linear dimensionality reduction techniques. However, we identified some limitations of the use of nonlinear manifold learning for collaborative filtering. In this paper, they will be considered.

The rest of the paper is organized as follows. The next section gives a brief overview of dimension reduction algorithms for collaborative filtering. Section 3 describes the selected two nonlinear techniques for dimensionality reduction and presents some of the challenges of these two nonlinear techniques for collaborative filtering. Section 4 delineates the assessment of non linear dimensionality reduction in the CF context. The final section provides some concluding remarks and future research directions.

2. Background

Recommender systems emerged as an independent research area in the mid-1990s when researchers started focusing on recommendation problems that explicitly rely on the ratings structure. In its most common formulation, the recommendation problem is reduced to the problem of estimating ratings for the items that have not been seen by a user. This estimation is usually based on the ratings given by this user to some other items. Once we can estimate ratings for the yet unrated items, we can recommend to the user the item(s) with the highest estimated rating(s).

Collaborative filtering utilizes the known preferences of a group of users to predict the unknown preference of a new user. However, the existing CF techniques have the drawback that requires the entire existing data be maintained and analyzed repeatedly whenever new user ratings are added. To avoid such a problem and improve computational efficiency of the existing CF techniques, a new approach called Eigentaste was proposed based on the principal component analysis (PCA). However, Eigentaste requires that each user rate every item in the so called gauge set for executing PCA, which may not be always feasible in practice [4].

The next study was an iterative PCA approach in which no gauge set is required [5]. The developed approach and Eigentaste, combined with two clustering methods, are compared in terms of the mean absolute error (MAE) of prediction using three real data sets. Computational results indicate that

the prediction accuracy of the proposed approach does not deteriorate even without a gauge set, and therefore, the proposed approach may be considered as a useful alternative when it is neither possible nor practical to define a gauge set. The iterative PCA approach using SVD takes a considerable amount of time and space to estimate a new user's missing ratings. Moreover, SVD based algorithms suffer from a serious drawback - the offline SVD decomposition step is computationally very expensive.

To alleviate this problem, two SVD update methods, the Zha and Simon [6] and the folding-in [2], are considered as possible alternatives. These alternatives are compared in terms of both the MAE and computational time using a real data set. The experimental results show that the SVD update method by Zha and Simon is better than the folding-in method.

Since the ratings data for CF reflecting the many-sided interests of many users could have nonlinear dependencies, a dimension reduction technique based on nonlinear PCA is developed. The proposed method can update the local mean in each cluster whenever a new user enters the system, and then, the updated local mean is used for the next new users. The experimental results reveal that this nonlinear PCA approach has a decreasing MAE as the number of new users increase, due to the updating of the local mean. This is a desirable result since it implies that the developed approach can be applied, even if a large number of new users enter the system continuously. Finally, the three approaches developed are compared in terms of the MAE and computational time. From the experiments in which the prediction accuracy and computational time are used as the comparative criterion, it is concluded that the method by Zha and Simon is the best as a dimension reduction technique for CF.

An attempt to apply nonlinear dimensionality reduction in recommender systems was described in some papers [7], [8], [9]. They observed that manifold alignment using non-linear dimensionality reduction has the promise of an effective supervised learning technique for the purposes of cross system personalization. When a large number of users cross over from one system to another, carrying their user profiles with them, a mapping between the user profiles of the two systems can be discovered. The key idea is to embed user profiles from different systems in low-dimensional manifolds such that profiles known to be in correspondence (i.e. profiles of the same user) are mapped to the same point. This means the manifolds will be aligned at correspondence points. A simple NLDR to a manifold performs better than popular voting. The predicted votes become more accurate as more users cross over and their profiles are aligned. But, the predictions are worse than SVD or Pearson's correlation based algorithm, which serve as the gold standard.

Another possible approach to produce low-dimensional presentation of the data is the use of domain-specific classification information to divide original user-item rating matrix into several low-dimensional dense user-item rating matrices [9]. Each item can be assigned to one or more classifications. For example, in the domain of movies, each movie can be classified according to the attribute "genre" of each item (the values of genre include Action, Adventure, Drama, and so on). In the domain of books, an attribute "category" of items is used to classify books.

3. The Use of Nonlinear Techniques for Dimensionality Reduction in Recommender Systems

In this section, we first review the non linear dimensionality reduction methods. More specifically, we focus on the selected two

nonlinear techniques for dimensionality reduction and present some of the challenges of these two nonlinear techniques for collaborative filtering.

3.1 Overview of Non Linear Dimensionality Reduction Methods

Assume we have dataset represented in a $n \times D$ matrix X consisting of n datavectors x_i ($i \in \{1, 2, \ldots, n\}$) with dimensionality D. Assume further that this dataset has intrinsic dimensionality d (where $d < D$, and often $d << D$). Here, in mathematical terms, intrinsic dimensionality means that the points in dataset X are lying on or near a manifold with dimensionality d that is embedded in the D-dimensional space. Dimensionality reduction techniques transform dataset X with dimensionality D into a new dataset Y with dimensionality d, while retaining the geometry of the data as much as possible. We shall assume that different y_i do not lie randomly in R^d, but approximately on a manifold, which is denoted by M. The manifold may simply be a hyperplane, or it can be more complicated. An example of a "curved" manifold with the data points lying on it can be seen in Figure 1. In general, neither the geometry of the data manifold, nor the intrinsic dimensionality d of the dataset X is known. Therefore, dimensionality reduction is an ill-posed problem that can only be solved by assuming certain properties of the data (such as its intrinsic dimensionality).

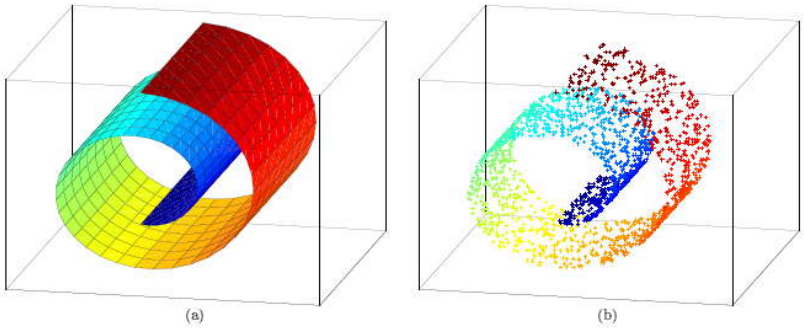

Figure 1. An example of a manifold. This example is usually known as the "Swiss roll". (a) Surface of the manifold. (b) Data points lying on the manifold.

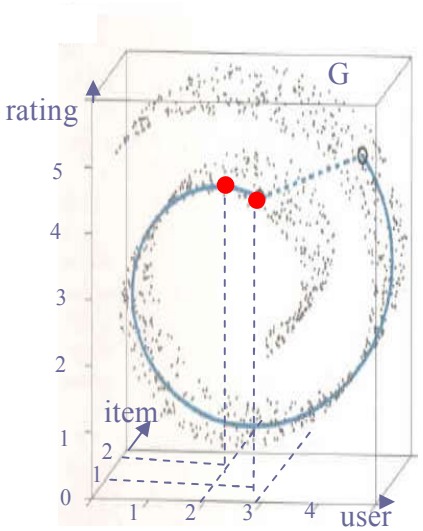

Figure 2. An example of a "curved" manifold

Under the assumption, that manifold in Figure 2 represents ratings given by the different users to the different items, we are trying to capture the relationships among pairs of customer based on ratings of products. The main distinction between techniques for dimensionality reduction is the distinction between linear and nonlinear techniques. Linear techniques assume that the data lie on or near a linear subspace of the high-dimensional space. Nonlinear techniques for dimensionality reduction do not rely on the linearity assumption as a result of which more complex embeddings of the data in the high-dimensional space can be identified. Most nonlinear techniques for dimensionality reduction have been proposed more recently and are therefore less well studied. In this section, we discuss two nonlinear manifold learning techniques recently proposed (Isomap [10] and Locally Linear Embedding (LLE) [11]).

3.2 Isomap

Multidimensional scaling has proven to be successful in many applications, but it suffers from the fact that it is based on Euclidean distances, and does not take into account the distribution of the neighboring datapoints. If the high-dimensional data lies on or near a curved manifold, such as in the Swiss roll dataset, MDS might consider two datapoints as near points, whereas their distance over the manifold is much larger than the typical interpoint distance. Isomap [12] is a technique that resolves this problem by attempting to preserve pairwise geodesic (or curvelinear) distances between datapoints. Geodesic distance is the distance between two points measured over the manifold.

In Isomap [12], the geodesic distances between the datapoints x_i ($i \in \{1, 2, \ldots, n\}$) are computed by constructing a neighborhood graph G, in which every datapoint x_i is connected with its k nearest neighbors x_{ij} ($j \in \{1, 2, \ldots, n\}$) in the dataset X. The shortest path between two points in the graph forms a good (over)estimate of the geodesic distance between these two points, and can easily be computed using Dijkstra's or Floyd's shortest-path algorithm [13,14]. The geodesic distances between all datapoints in X are computed, thereby forming a pairwise geodesic distance matrix. The low-dimensional representations y_i of the datapoints x_i in the low-dimensional space Y are computed by applying multidimensional scaling on the resulting distance matrix. An important weakness of the Isomap algorithm is its topological instability [15]. Isomap may construct erroneous connections in the neighborhood graph G. In the case of user-item matrix, $i \times j$ matrix M consisting of ratings r given by the i users to the j items, user's and item's places are randomly posited in matrix. Therefore, the rating points lie randomly on a manifold. There is no reason to respect this shape trying to find similar users or similar items. It is possible to identify two users which rate different items, as neighbours on this manifold (dots marked in the Figure 2). It doesn't mean that these users rate similarly.

Such short-circuiting [16] can severely impair the performance of Isomap. Several approaches have been proposed to overcome the problem of short-circuiting, e.g., by removing datapoints with large total flows in the shortest path-algorithm [17] or by removing nearest neighbors that violate local linearity of the neighborhood graph [18]. A second weakness is that Isomap may suffer from 'holes' in the manifold. This problem can be dealt with by tearing manifolds with holes [19]. A third weakness of Isomap is that it can fail if the manifold is nonconvex [20].

3.3 LLE

Local Linear Embedding (LLE) [21] is a local technique for dimensionality reduction that is similar to Isomap i.e. it constructs a

graph representation of the datapoints. In contrast to Isomap, it attempts to preserve solely local properties of the data. The preservation of local properties allows successful embedding of nonconvex manifolds. In LLE, the local properties of the data manifold are constructed by writing the datapoints as a linear combination of their nearest neighbors. In the low-dimensional representation of the data, LLE attempts to retain the reconstruction weights in the linear combinations as good as possible.

LLE describes the local properties of the manifold around a datapoint x_i by writing the datapoint as a linear combination W_i (the so-called reconstruction weights) of its k nearest neighbors x_{ij}. Hence, LLE fits a hyperplane through the datapoint x_i and its nearest neighbors, thereby assuming that the manifold is locally linear. The local linearity assumption implies that the reconstruction weights W_i of the datapoints x_i are invariant to translation, rotation, and rescaling. Because of the invariance to these transformations, any linear mapping of the hyperplane to a space of lower dimensionality preserves the reconstruction weights in the space of lower dimensionality. In other words, if the low-dimensional data representation preserves the local geometry of the manifold, the reconstruction weights W_i that reconstruct datapoint x_i from its neighbors in the high-dimensional data representation also reconstruct datapoint y_i from its neighbors in the low-dimensional data representation. This is done by choosing d-dimensional coordinates of Y to minimize the embedding cost function:

$$\phi(Y) = \sum(y_i - \sum_{j=1}^{k} w_{ij} y_{ij})^2$$

It can be shown that the coordinates of the lowdimensional representations y_i that minimize this cost function can be found by computing the eigenvectors corresponding to the smallest d nonzero eigenvalues of the inproduct $(I - W)^T (I - W)$. In this formula, I is the $n \times n$ identity matrix. LLE tends to collapse large portions of the data onto a single point. In [22], it is claimed that LLE performs worse than Isomap, because LLE has difficulties when confronted with manifolds that contain holes. Some authors also make negative conclusion that LLE is only useful for small numbers of dimensions [12]. A possible explanation is that the practical data includes a large number of intrinsic features and have high curvature both in the observation space and in the embedded space, whereas present manifold learning methods strongly depends on the selection of parameters.

4. Assessment of Non Linear Dimensionality Reduction in the CF context

The tremendous growth of customers and products in E-commerce domain has motivated explorations of linear and nonlinear techniques for dimensionality reduction, both as a compressed representation of the data and as a basis for recommendations via regression. The main drawback of dimensionality reduction is the possibility of information loss. When done poorly, dimensionality reduction can discard useful instead of irrelevant information.

Linear regression models generally have lower sample complexity per parameter than nonlinear and nonparametric models. The underlying assumption for nonlinear dimensionality reduction is that the data points do not lie randomly in the high-dimensional space; rather, there is a certain structure in the locations of the data points that can be exploited, and the useful information in high dimensional data can be summarized by a small number of attributes. Local approaches like LLE attempt to preserve the local geometry of the data; essentially, they seek to map nearby points on the manifold to nearby points in the low-

dimensional representation. Global approaches like Isomap attempt to preserve geometry at all scales, mapping nearby points on the manifold to nearby points in low-dimensional space, and faraway points to faraway points.

In the case of user-item matrix, matrix consisting of ratings given by users to items, user's and item's places are randomly positioned in matrix. Therefore, the rating points lie randomly on a manifold. There is no reason to respect this shape trying to find similar users or similar items. It is possible that nonlinear manifold learning techniques identify two users which rate different items, as neighbors on this manifold. It doesn't mean that these users rate similarly.

5. Conclusions and Future Work

The current generation of recommender systems requires further improvements to make recommendation methods more effective in a broader range of applications. In this paper, we reviewed various limitations of the use of nonlinear manifold learning for collaborative filtering recommendation methods.

We hope that the issues presented in this paper will advance the discussion in the recommender systems community about the next generation of recommendation technologies.

Regarding future work, comparisons with further nonlinear methods need to be conducted. A comprehensive study, mentioned above, would be beneficial for the collaborative filtering as well as for scientists, who have to deal with high-dimensional data.

References

[1] G. Linden, B. Smith, and J. York. *Amazon.com recommendations: Item-to-item collaborative filtering*. IEEE Internet Computing, 2003, pp. 76–80

[2] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl: *Incremental Singular Value Decomposition Algorithms for Highly Scalable Recommender Systems*, Department of Computer Science and Engineering University of Minnesota, In ACM WebKDD 2001 Web Mining for E-Commerce Workshop, 2001.

[3] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. *Application of dimensionality reduction in recommender systems–a case study*. In ACM WebKDD 2000 Web Mining for E-Commerce *Workshop*, 2000.

[4] Goldberg, K., Roeder, T., Gupta, D., & Perkins, C. *Eigentaste: A constant time collaborative filtering algorithm.* Information Retrieval *4(2)*, 2001, pp. 133–151

[5] D. Kim and B. Yum. *Collaborative filtering based on iterative principal component analysis*. Expert Systems with Applications, 28(4), 2005, pp. 823-830

[6] H. Zha and H. D. Simon. *On updating problems in latent semantic indexing*. SIAM Journal on Scientific Computing, 21(2), 1999, pp. 782–791

[7] B. Mehta, C. Niederee, and A. Stewart. *Towards cross-system personalization*. In UAHCI, 2005.

[8] B. Mehta, C. Nieder´ee, A. Stewart, M. Degemmis, P. Lops, and G. Semeraro. *Ontologically-enriched unified user modeling for cross-system personalization. In User Modeling*, 2005, pp. 119–123

[9] T. Gao, C. Xing, Y. Zhao: *An Effective Algorithm for Dimensional Reduction in Collaborative Filtering,* ICADL 2007, Hanoi, Vietnam, December 10-13., Springer, 2007.

[10] J. B. Tenenbaum, V. de Silva, and J. C. Langford, *A global geometric framework for nonlinear dimensionality*

reduction, Science 290, 2000, pp. 2319-2323

[11] S. Roweis and L. Saul *Nonlinear dimensionality reduction by locally linear embedding*, Science 290, 2000, pp. 2323-2326

[12] L. Teng, H. Li, X. Fu, W. Chen, and I.-F. Shen. *Dimension reduction of microarray data based on local tangent space alignment*. In Proceedings of the 4th IEEE International Conference on Cognitive Informatics, 2005, pp. 154–159

[13] E.W. Dijkstra. *A note on two problems in connexion with graphs.* Numerische Mathematik, 1, 1959, pp. 269–271

[14] R.W. Floyd. *Algorithm 97: Shortest path.* Communications of the ACM, 5(6), 1962, 345p

[15] M. Balasubramanian and E.L. Schwartz. *The Isomap algorithm and topological stability*. Science, 295(5552):7, 2002.

[16] H. Li, L. Teng, W. Chen, and I.-F. Shen. *Supervised learning on local tangent space. In Lecture Notes on Computer Science*, volume 3496, Berlin, Germany, 2005, Springer Verlag, pp. 546–551

[17] H. Choi and S. Choi. *Robust kernel Isomap. Pattern Recognition*, 40(3), 2007, pp. 853–862

[18] F. Sha and L.K. Saul. *Analysis and extension of spectral methods for nonlinear dimensionality reduction.* In Proceedings of the 22nd International Conference on Machine Learning, 2005, pp. 785–792

[19] B.L. Betechuoh and T. Marwalaand T. Tettey. *Autoencoder networks for HIV classification*. Current Science, 91(11), 2006, pp. 1467–1473

[20] M.E. Tipping. *Sparse kernel principal component analysis.* In Advances in Neural Information Processing Systems, volume 13, Cambridge, MA, USA, 2000, The MIT Press. pp 633–639

[21] L.K. Saul, K.Q. Weinberger, J.H. Ham, F. Sha, and D.D. Lee. *Spectral methods for dimensionality reduction. In Semisupervised Learning*, Cambridge, MA, USA, 2006, The MIT Press.

[22] N. Mekuz and J.K. Tsotsos. *Parameterless Isomap with adaptive neighborhood selection. In Proceedings of the 28th DAGM Symposium*, Berlin, Germany, 2006, Springer., pp. 364–373