ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Proximity learning for non-standard big data

Frank-Michael Schleif

The University of Birmingham, School of Computer Science,
Edgbaston Birmingham B15 2TT, United Kingdom.

Abstract. Huge and heterogeneous data sets, e.g. in the life science domain, are challenging for most data analysis algorithms. State of the art approaches do often not scale to larger problems or are inaccessible due to the variety of the data formats. A flexible and effective method to analyze a large variety of data formats is given by proximity learning methods, currently limited to medium size, *metric* data. Here we discuss novel strategies to open relational methods for non-standard data at large scale, applied to a very large protein sequence database.

1 Introduction

In many application areas such as bioinformatics, physics, or the web, electronic data is getting larger and more complex in size and representation. The challenges are manifold often identified by volume, variety, velocity and variability accompanied by veracity and summarized in the term *big data*. In the following we show and link some recently developed techniques to address some of the *v*'s by using well established mathematical and computational intelligence models, like kernel machines or prototype based learning algorithms [8, 11]. The analysis of large volume data sets is an always challenging task but it gets even more complicated if the data are in a non-standard format, e.g. non-vectorial and non-metric. We restrict our analysis to metric and non-metric similarities but the approach can also be used for (non-)metric dissimilarities as detailed in [9]. In general the models are define by a combination of those proximities from a training set, see e.g. [3]. In this way it is also possible to use own user defined metric proximity functions such that very flexible data representation become possible. Many of these models permit a clear communication of the model decision process, because the decision function is based on identifiable training points. New data points can be mapped to these models by calculating proximities between the query object and a small number of proximities, used in the model [10].

Most of these approaches are however limited to problems at moderate volume, e.g. multiple thousand objects and require either on the fly or on block calculations of a large number of pairwise proximities, being unattractive for large to very large problems. Further a need of *metric* proximities cancel out a large number of domain specific, non-metric proximity scores, like the non-metric *Smith-Waterman* [4] alignment score for protein sequences. The goal of the article is to discuss strategies which extend proximity learning methods to the *large scale* also for non-metric proximities. With large we refer to multiple hundred thousand to millions of objects and a corresponding squared number of considered pairwise proximities. First we briefly review some recent achievements in this line followed by an extension of a recent approach to *non-metric similarity data*[1]. Experimental results for a large protein sequence database show the effectiveness of the proposed method.

[1] Due to lack of space we do not discuss dissimilarities which are addressed in more detail in [9]

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

2 Proximity learning for non-standard data

Many classical machine learning techniques, have been proposed for Euclidean vectorial data. However, modern data are often associated to dedicated structures making a representation in terms of Euclidean vectors difficult: biological sequence data, text files, graphs, or time series [2]. These data are inherently compositional and a feature representation leads to information loss. As an alternative, a dedicated proximity measure such as pairwise alignment, or kernels for structures can be used as the interface to the data. Native methods for the analysis of proximity data have been proposed in [8, 3], but are widely based on non-convex optimization schemes and with quadratic to linear memory and runtime complexity. Since most analysis methods rely on *metric* input data of the underlying similarities, different preprocessing approaches have been analyzed to correct non-metric or non psd similarity matrices [2], typically based on eigenvalue corrections. The transformation between the different representations as well as the correction approaches have typically quadratic or cubic costs. In [10] the author proposed a method for the transformation between the different representations including such an eigenvalue correction with linear costs for moderately large datasets based on the Nyström approximation [12]. This transformation represents a given matrix by a small number of so called landmark points and their relation to the remaining data points, as detailed later on. As an inherent step a quadratic matrix of these landmark proximities is used. This can become very costly for larger datasets, where either the accuracy of the approximation suffers, due to a small number of landmarks, or the approximation costs raise if a sufficiently large number of landmarks is used. In [6] a new approach for the calculation of the Nyström approximation for *psd* matrices was proposed using a random projection technique. As shown in [6] this strategy is very effective to keep high accuracy of the matrix approximations also for very large data sets. Motivated by these promising results we will derive an extension of our former approach published in [10] using similar strategies. Our objective is to provide a method of linear complexity to transform similarities into dissimilarities and vice versa including potential eigenvalue corrections to make the problem psd. Now we will first briefly review previous work of the author proposed in [10] followed by a discussion of the subspace Nyström approximation proposed in [6], both linked to each other later on.

2.1 Approximation of proximities at large scale

For kernel methods and more recently for prototype based learning the usage of the Nyström approximation is a well known technique to approximate proximity matrices to obtain effective learning algorithms [12, 3]. A kernel matrix is approximated by

$$\tilde{\mathbf{K}} = \mathbf{K}_{m,N}^{\top} \mathbf{K}_{m,m}^{-1} \mathbf{K}_{m,N}. \tag{1}$$

where $\mathbf{K}_{m,m}^{-1}$ denoting the Moore-Penrose pseudoinverse of the landmark matrix and $\mathbf{K}_{m,N}$ is the matrix of the landmark similarities to all other points. This approximation is exact, if $\mathbf{K}_{m,m}$ has the same rank as $\mathbf{K}$ (for details see e.g. [12]).

A benefit of the Nyström technique is that it can be decided priorly which linear parts of the dissimilarity matrix will be used in training. Therefore, it is sufficient to *compute only a linear part of the full proximity matrix* to use these methods.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

2.2 Transformations and corrections of proximities with linear costs

For *metric* similarity data, kernel methods can be applied directly, or in case of large N, the Nyström approximation can be used. For non-metric similarities an eigenvalue correction has to be done first, which for the full matrix would again be very costly. The Nyström approximation can again decrease computational costs dramatically. Since we now can apply the approximation on an arbitrary symmetric matrix, we can make the correction afterward. To correct an already approximated similarity matrix $\hat{\mathbf{S}}$ it is sufficient to correct the eigenvalues of $\mathbf{S}_{m,m}$. Altogether we get $\mathcal{O}(m^2N)$ complexity.

We can write for the approximated matrix $\hat{\mathbf{S}}$ its eigenvalue decomposition as $\hat{\mathbf{S}} = \mathbf{S}_{N,m}\mathbf{S}_{m,m}^{-1}\mathbf{S}_{N,m}^{\top} = \mathbf{S}_{N,m}\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^{\top}\mathbf{S}_{N,m}^{\top}$, where we can correct the eigenvalues $\mathbf{\Lambda}$ by some technique as discussed e.g. in [2] to $\mathbf{\Lambda}^*$. In general flipping (flip negative eigenvalues) or clipping (removing negative eigenvalues) are used. The corrected approximated matrix $\hat{\mathbf{S}}^*$ is then simply

$$\hat{\mathbf{S}}^* = \mathbf{S}_{N,m}\mathbf{U}\left(\mathbf{\Lambda}^*\right)^{-1}\mathbf{U}^{\top}\mathbf{S}_{N,m}^{\top}. \tag{2}$$

If it is desirable to work with the corrected dissimilarities, then we should note, that it is possible to transform the similarity matrix $\mathbf{S}$ to a dissimilarity matrix $\mathbf{D}$: $D_{ij}^2 = S_{ii} + S_{jj} - 2S_{ij}$. For the generalization to new unseen test data an out of sample extension is needed. This might be a problem for the techniques dealing with (dis)similarities. If the matrices are corrected, we need to correct the new (dis)similarities as well to get consistent results. Fortunately, it is quite easy in the Nyström framework. By examining Eq. (2) we see, that we simply need to extend the matrices $\mathbf{S}_{N,m}$, respectively, by uncorrected similarities between the new points and the landmarks to obtain the full approximated and *corrected* similarity matrices, which then can be used by the algorithms to compute the out of sample extension. A detailed discussion with experiments is available in [10].

3 Very large scale Nyström approximation for general proximities

In [6] a new approach for the calculation of the Nyström approximation for large *psd* matrices was proposed, we will denote this approach as LSNA and our proposal as extended LSNA (e-LSNA). As one can see in Eq. (1) the Nyström approximation is based on the calculation of a pseudo-inverse of a matrix based on m rows and m columns. The *optimal* landmarks specifying these columns and rows are the cluster centers of the considered data set, which are hard to identify in advance. Accordingly, m is often chosen to be sufficiently large such that the landmarks are likely to cover enough information of the data distribution. For large data sets, containing e.g. million of points the number of landmarks is also getting large e.g. $m = 10000$ and the calculation of the pseudo-inverse may dominate the remaining calculation costs due to the cubic complexity. The idea, presented in [6] is to use a randomized singular value decomposition (SVD) [5] on the landmark matrix to obtain an accurate $m \times m$ matrix in the Nyström approximation at low costs. Note that in the randomized SVD algorithm [5] arbitrary, squared matrices can be used, a psd assumption is not necessary.

Thereby the data are represented on a lower dimensional subspace e.g. in k dimensions with $k \ll m$ such that the obtained singular value matrix L can be inverted with

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

low costs of $O(k^3)$. The final Nyström approximation of the original proximity matrix can be obtained by subsequent matrix multiplications leading to a similar formulation as before. The basic algorithm taken from [6] is shown in Alg. 1. Here, p is an over-sampling parameter (typically set to 5 or 10) such that the rank of Q is slightly larger than the desired rank(k), and q is the number of steps of a power iteration (typically set to 1 or 2) which is used to speed up the decay of the singular values of W [6].

Algorithm 1 Large scale Nyström approximation [6]

1: **init:** psd matrix $K \in \mathbb{R}^{N \times N}$, #landmarks m, rank k, over-sampling p, power q
2: **Output:** $\hat{K}$, an approximation of K
3: $C \leftarrow m$ columns of K sampled uniformly at random without replacement
4: $W \leftarrow m \times m$ landmark matrix
5: $[\tilde{U}, \hat{\Lambda}] \leftarrow ranksvd(W, k, p, q)$ using Alg. proposed in [5], $U \leftarrow C\tilde{U}\Lambda^{-1}$
6: $\hat{K} \leftarrow \left(\sqrt{\frac{m}{N}}U\right)\left(\frac{m}{N}\Lambda\right)\left(\sqrt{\frac{m}{N}}U^\top\right)$

If the given proximity data are psd similarities, algorithm 1 can be used directly. For non-psd similarities the SVD used in algorithm 1 implicitly flips negative eigenvalues. Due to the random projection step it may obviously also happen that smaller absolute eigenvalues are removed. While this appears to be a nice feature it is not always clear if flipping is a good strategy. If the data are metric dissimilarities the approach can still be used with some modification and the same complexity, see [9] for further details.

3.1 Non-metric similarities

The LSNA approach performs an implicit flipping of negative eigenvalues in the SVD step, to get more control about the handling of negative eigenvalues we will introduce an explicit step to correct the eigenvalue but in the low dimensional projection space with low computational costs shown in Alg. 2 and 3.

Algorithm 2 Randomized SVD with eigenvalue correction

1: **init:** $m \times m$ matrix W, scalars k, p, q
2: **Output:** U, Λ, V
3: $\Omega \leftarrow m \times (k + p)$ standard Gaussian random matrix
4: $Z \leftarrow W\Omega$, $Y \leftarrow W^{q-1}Z$,
5: find orthonormal Q such that $Y = QQ^\top Y$
6: $B(Q^\top Q) = Q^\top Z$
7: $[E_b, V_b] = eig(B)$, $V_b^* \leftarrow$ flip—clip—shift(V_b), $B^* = E_b \cdot V_b^* \cdot E_b^\top$
8: $[V, L, V'] = svd(B^*)$
9: $U \leftarrow QV'$

The new formulation accounts for an explicit eigenvalue correction in Algorithm 2. Note that **both** unitary matrices V and V' are needed, to get valid reconstructions. To make the out of sample extension more obvious, it is also convenient to modify the reconstruction of the proximity matrix in Alg. 3, line 6. Now it can be directly seen how to extended the matrix K by l items. One only needs to calculate the corresponding m similarities to the m landmarks which can become part of an extended matrix C.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Algorithm 3 Large scale Nyström approximation with eigenvalue correction

1: **init:** (non-)psd matrix $K \in \mathbb{R}^{N \times N}$, #landmarks m, rank k, p, q
2: **Output:** $\hat{K}$, an approximation of K
3: $C \leftarrow m$ columns of K sampled uniformly at random without replacement
4: $W \leftarrow m \times m$ landmark matrix
5: $[\tilde{U}, \hat{\Lambda}, V] \leftarrow ranksvd(W, k, p, q)$ using Alg. 2
6: $\hat{K} \leftarrow \left(\sqrt{\frac{m}{N}}C\right)\left(\frac{m}{N}V(\tilde{U}\Lambda^{-1})^\top\right)\left(\sqrt{\frac{m}{N}}C^\top\right)$

Equation (2) can be modified to integrate Algorithm 3 straight forward by replacing $\mathbf{S}_{N,m}$ with $\sqrt{\frac{m}{N}}C$ and $\mathbf{U}\left(\mathbf{\Lambda}^*\right)^{-1}\mathbf{U}^\top$ with $\left(\frac{m}{N}V(\tilde{U}\Lambda^{-1})^\top\right)$.

The complete costs of this operations are $O(Nmk + k^3)$ which has still the same complexity as the original LSNA algorithm. We are now able to process potentially non-metric similarities at large scale. A reconstruction of the original matrices, without an eigenvalue correction, can be used to analyze the approximation error by means of the frobenius norm and the preservation of ranks. An analysis of this type was done in [10] and [9] for this type of algorithm. Empirically, we found that the observed error is small as long as the rank of the data space is sufficiently preserved by the landmark matrix W, further the parameter k in the SVD should not exceed the intrinsic rank of the matrix W to avoid noise amplification.

4 Explore the large protein space by kernel methods

In our experiments we consider the SwissProt protein database [1] of 11/2010, restricted to ProSite labeled sequences with at least 100 entries per label. We obtain 521 ProSite labels and 265.166 sequences which are compared by the Smith-Waterman alignment algorithm as provided in [7]. The obtained similarity scores are symmetric but non-metric, accordingly standard kernel methods can not be used directly in a valid form. Our objective is now to use a kernel classifier on this type of data. For new sequences it would be very desirable to find those ProSite labels which are most typical to the new unknown sequence, e.g. to judge its biochemical properties. Also a low-dimensional visualization by a laplacian eigenmap is shown, using ProSite labels as colors.

Although for kernel classifiers sequence data can be handled by dedicated kernels we would like to use the gold-standard domain measures. A standard approach is costly in calculating the whole similarity matrix and it would be basically impossible to get an eigenvalue correction in a reasonable time. Modern kernel classifiers like the Core-Vector Machine (CVM)[11] do not need to evaluate all the kernel similarities but our similarities are non-metric and an accurate online eigenvalue correction is not available.

However we can use our presented approach approximating the score matrix as well as performing an eigenvalue correction. To do this we specify 3354 landmarks, randomly taken from the 521 most prominent classes. The remaining parameters are $k = 400$,$p = 10$ and $q = 2$. Accordingly we need only around 7 GB to store the matrix C and W and some days to calculate the scores. The approximation and eigenvalue correction by the presented approach takes only some minutes. The obtained approximated

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Eigenvalue corr.	flip	clip
Prediction acc.	$76.85 \pm 0.36\%$	77.00 ± 0.31

Table 1: Crossvalidation results of the SwissProt data using flip or clip correction.

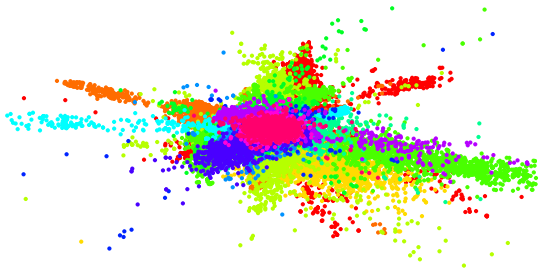

Fig. 1: Eigenmap of the SwissProt data (largest 21 ProSite classes).

and now positive semi definite similarity matrix is used by a one-class Core-Vector Machine in a 5 fold crossvalidation to generate a classification model with a good mean prediction accuracy see Table 1. An additional benefit of the CVM approach is that it naturally leads to very sparse models. Accordingly the out of sample extension to new sequences requires only few score calculations to the sequences of the training set.

Using the same encoded similarity matrix and the approach used in [6] we can now, having a psd similarity matrix, calculate a laplacian eigenmap of our dataset. To avoid clutter we show only a visualization of the 21 largest classes in Figure 1. This visualization can be used to analyze local relations of the different sequences.

5 Conclusion

We presented a new method for the analysis of large scale (non-metric) proximity matrices. The approach scales to multiple million objects and is only limited by the memory needed to store the matrix C. Our approach can be used to analyze large non-metric proximity matrices by means of relational learning approaches. In the experiments we have shown the effectiveness of the method for a large scale life science problem, but the approach is more generic applicable to a wide range of proximity data sets [2]

References

[1] B. Boeckmann. The swiss-prot protein knowledgebase and its suppl. trembl in 2003,. *Nucleic Acids Res.*, 31:365–370.
[2] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *JMLR*, 10:747–776, 2009.
[3] A. Gisbrecht, B. Mokbel, F.-M. Schleif, X. Zhu, and B. Hammer. Linear time relational prototype based learning. *Journal of Neural Systems*, 22(5), 2012.
[4] D. Gusfield. *Alg. on Strings, Trees, and Seq.: Comp. Sc. and Comp. Biology*. Cambridge University Press, 1997.
[5] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.
[6] M. Li, J. T. Kwok, and B.-L. Lu. Making large-scale nyström approximation possible. In *ICML*, pages 631–638, 2010.
[7] Mathworks Inc. Matlab 2011a. http://www.mathworks.com (21.11.2013), 2011.
[8] E. Pekalska and R. Duin. *The dissimilarity representation for pattern recognition*. World Scientific, 2005.
[9] F.-M. Schleif. Large scale Nyström approximation for non-metric similarity and dissimilarity data. Technical Report MLR-03-2013, ISSN:1865-3960 http://www.uni-leipzig.de/compint/mlr/mlr_03_2013.pdf, 2013.
[10] F.-M. Schleif and A. Gisbrecht. Data analysis of (non-)metric proximities at linear costs. In *SIMBAD 2013*, pages 59–74, 2013.
[11] I. W. Tsang, A. Kocsor, and J. T. Kwok. Simpler core vector machines with enclosing balls. In *ICML*, volume 227 of *ACM International Conference Proceeding Series*, pages 911–918. ACM, 2007.
[12] C. K. I. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *NIPS*, pages 682–688. MIT Press, 2000.

[2] **Acknowledgment**: I would like to thank Andrej Gisbrecht for prior work on Nyström approximation and Xibin Zhu for providing initial support with the SwissProt sequence database. This work was funded by a EU Marie Curie Intra-European Fellowship (PIEF-GA-2012-327791).