Manifold Learning and the Quantum Jensen-Shannon Divergence Kernel

Luca Rossi[1], Andrea Torsello[1], and Edwin R. Hancock[2]

[1] Department of Environmental Science, Informatics and Statistics,
Ca' Foscari University of Venice, Italy
{lurossi,torsello}@dsi.unive.it
[2] Department of Computer Science, University of York, YO10 5GH, UK
edwin.hancock@york.ac.uk

Abstract. The quantum Jensen-Shannon divergence kernel [1] was recently introduced in the context of unattributed graphs where it was shown to outperform several commonly used alternatives. In this paper, we study the separability properties of this kernel and we propose a way to compute a low-dimensional kernel embedding where the separation of the different classes is enhanced. The idea stems from the observation that the multidimensional scaling embeddings on this kernel show a strong horseshoe shape distribution, a pattern which is known to arise when long range distances are not estimated accurately. Here we propose to use Isomap to embed the graphs using only local distance information onto a new vectorial space with a higher class separability. The experimental evaluation shows the effectiveness of the proposed approach.

Keywords: Graph Kernels, Manifold Learning, Continuous-Time Quantum Walk, Quantum Jensen-Shannon Divergence.

1 Introduction

Graph-based representations have become increasingly popular due to their ability to characterize in a natural way a large number of systems [2, 3]. Unfortunately, our ability to analyse this wealth of data is severely limited by the restrictions posed by standard pattern recognition techniques, which usually require the graphs to be first embedded into a vectorial space, a procedure which is far from being trivial. Kernel methods [4] provide a neat way to shift the problem from that of finding an embedding to that of defining a positive semidefinite kernel. In fact, once we define a positive semidefinite kernel $k : X \times X \to \mathbb{R}$ on a set X, there exists a map $\phi : X \to H$ into a Hilbert space H, such that $k(x, y) = \phi(x)^\top \phi(y)$ for all $x, y \in X$. Thus, any algorithm can be formulated in terms of the data by implicitily mapping them to H via the well-known kernel trick. As a consequence, we are now faced with the problem of defining a positive semidefinite kernel on graphs rather than computing an embedding. However, due to the rich expressiveness of graphs, this task has also proven to be difficult.

Many different graph kernels have been proposed in the literature [5–7], which are generally instances of the family of R-convolution kernels introduced by

R. Wilson et al. (Eds.): CAIP 2013, Part I, LNCS 8047, pp. 62–69, 2013.

Haussler [8]. The fundamental idea is that of decomposing two discrete objects them and comparing some simpler substructures. For example, Gärtner et al. [5] propose to count the number of common random walks between two graphs, while Borgwardt and Kriegel [6] measure the similarity based on the shortest paths in the graphs. Shervashidze et al. [7], on the other hand, count the number of graphlets, i.e. subgraphs with k nodes. Recently, Rossi et. al [1] introduced a novel kernel where the graph structure is probed through the evolution of a continuous-time quantum walk [9]. The idea underpinning their method is that the interference effects introduced by the quantum walk seem to be enhanced by the presence of symmetrical motifs in the graph [10, 11]. To this end, they define a walk onto a new structure that is maximally symmetric when the original graphs are isomorphic. Finally, the kernel is defined as the quantum Jensen-Shannon divergence [12] between the density operators [13] associated with the walks.

In this paper, we study the separability properties of the QJSD kernel and we apply standard manifold learning techniques [14, 15] on the kernel embedding to map the data onto a low-dimensional space where the different classes can exhibit a better linear separation. The idea stems from the observation that the multidimensional scaling embeddings of the QJSD kernel show the so-called *horseshoe effect* [16]. This particular behaviour is known to arise when long range distances are not estimated accurately, and it implies that the data lie on a non-linear manifold. This is no surprise, since Emms et. al [10] have shown that the continuous-time quantum walk underestimates the commute time related to the classical random walk. For this reason, it is natural to investigate the impact of the locality of distance information on the performance of the QJSD kernel. Given a set of graphs, we propose to use Isomap [14] to embed the graphs onto a low-dimensional vectorial space, and we compute the separability of the graph classes as the distance information varies from local to global. Moreover, we perform the same analysis on a set of alternative graph kernels commonly found in the literature [5–7]. Experiments on several standard datasets demonstrate that the Isomap embedding shows a higher separability of the classes.

The remainder of this paper is organized as follows: Section 2 introduces some basic quantum mechanical terminology, while Section 3 reviews the QJSD kernel. Section 4 illustrates the experimental results and the conclusions are presented in Section 5.

2 Quantum Mechanical Background

Quantum walks are the quantum analogue of classical random walks. In this paper we consider only continuous-time quantum walks, as first introduced by Farhi and Gutmann in [9]. Given a graph $G = (V, E)$, the state space of the continuous-time quantum walk defined on G is the set of the vertices V of the graph. Unlike the classical case, where the evolution of the walk is governed by a stochastic matrix (i.e. a matrix whose columns sum to unity), in the quantum case the dynamics of the walker is governed by a complex unitary matrix i.e.,

a matrix that multiplied by its conjugate transpose yields the identity matrix. Hence, the evolution of the quantum walk is reversible, which implies that quantum walks are non-ergodic and do not possess a limiting distribution. Using Dirac notation, we denote the basis state corresponding to the walk being at vertex $u \in V$ as $|u\rangle$. A general state of the walk is a complex linear combination of the basis states, such that the state of the walk at time t is defined as

$$|\psi_t\rangle = \sum_{u \in V} \alpha_u(t) \, |u\rangle \tag{1}$$

where the amplitude $\alpha_u(t) \in \mathbb{C}$ and $|\psi_t\rangle \in \mathbb{C}^{|V|}$ are both complex.

At each instant in time the probability of the walker being at a particular vertex of the graph is given by the square of the norm of the amplitude of the relative state. More formally, let X^t be a random variable giving the location of the walker at time t. Then the probability of the walker being at the vertex u at time t is given by

$$\Pr(X^t = u) = \alpha_u(t)\alpha_u^*(t) \tag{2}$$

where $\alpha_u^*(t)$ is the complex conjugate of $\alpha_u(t)$. Moreover $\alpha_u(t)\alpha_u^*(t) \in [0,1]$, for all $u \in V$, $t \in \mathbb{R}^+$, and in a closed system $\sum_{u \in V} \alpha_u(t)\alpha_u^*(t) = 1$.

The evolution of the walk is governed by Schrödinger equation, where we take the Hamiltonian of the system to be the graph adjacency matrix A, which yields

$$\frac{d}{dt} |\psi_t\rangle = -iA \, |\psi_t\rangle \tag{3}$$

Given an initial state $|\psi_0\rangle$, we can solve Equation 3 to determine the state vector at time t

$$|\psi_t\rangle = e^{-iAt} \, |\psi_0\rangle = \Phi e^{-i\Lambda t} \Phi^\top \, |\psi_0\rangle \ , \tag{4}$$

where $A = \Phi\Lambda\Phi^\top$ is the spectral decomposition of the adjacency matrix.

Consider a quantum system that can be in a number of states $|\psi_i\rangle$ each with probability p_i. The system is said to be in the ensemble of (pure) states $\{|\psi_i\rangle \, , p_i\}$. The density operator (or density matrix) of such a system is defined as

$$\rho = \sum_i p_i \, |\psi_i\rangle \, \langle\psi_i| \tag{5}$$

The Von Neumann entropy [13] of a density operator ρ is $H_N(\rho) = -Tr(\rho \log \rho) = -\sum_j \lambda_j \log \lambda_j$, where the λ_js are the eigenvalues of ρ.

With the Von Neumann entropy to hand, we can define the quantum Jensen-Shannon divergence between two density operators ρ and σ as

$$D_{JS}(\rho, \sigma) = H_N\Big(\frac{\rho + \sigma}{2}\Big) - \frac{1}{2}H_N(\rho) - \frac{1}{2}H_N(\sigma) \tag{6}$$

This quantity is always well defined, symmetric and negative definite [17]. It can also be shown that $D_{JS}(\rho, \sigma)$ is bounded, i.e., $0 \leq D_{JS}(\rho, \sigma) \leq 1$, with equality to 1 if and only if the states ρ and σ have support on orthogonal subspaces.

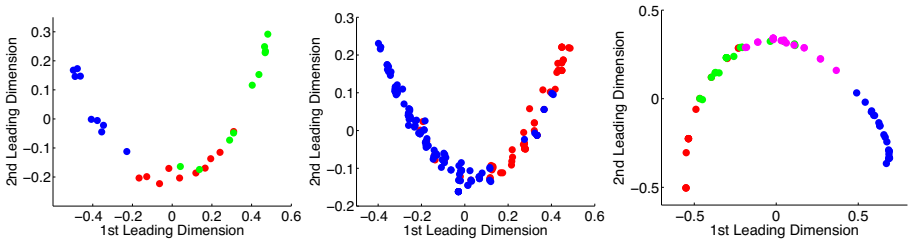

Fig. 1. The MDS embeddings from the QJSD kernel consistently show an horseshoe shape distribution of the points

3 The QJSD Kernel

Given two graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ we construct a new graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = V_1 \cup V_2$, $\mathcal{E} = E_1 \cup E_2 \cup E_{12}$, and $(u, v) \in E_{12}$ only if $u \in V_1$ and $v \in V_2$.With this new structure to hand, we define two continuous-time quantum walks $\left|\psi_t^-\right\rangle = \sum_{u \in V} \psi_{0u}^- \left|u\right\rangle$ and $\left|\psi_t^+\right\rangle = \sum_{u \in V} \psi_{0u}^+ \left|u\right\rangle$ on $\mathcal{G}$ with starting states

$$\psi_{0u}^- = \begin{cases} +\frac{d_u}{C} \text{ if } u \in G_1 \\ -\frac{d_u}{C} \text{ if } u \in G_2 \end{cases} \qquad \psi_{0u}^+ = \begin{cases} +\frac{d_u}{C} \text{ if } u \in G_1 \\ +\frac{d_u}{C} \text{ if } u \in G_2 \end{cases} \tag{7}$$

where d_u is the degree of the node u and C is the normalisation constant such that the probabilities sum to one.

We allow the two quantum walks evolve until a time T and we define the average density operators ρ_T and σ_T over this time as

$$\rho_T = \frac{1}{T} \int_0^T \left|\psi_t^-\right\rangle \left\langle\psi_t^-\right| \, \mathrm{d}t \qquad \sigma_T = \frac{1}{T} \int_0^T \left|\psi_t^+\right\rangle \left\langle\psi_t^+\right| \, \mathrm{d}t \tag{8}$$

In other words, we have defined two mixed systems with equal probability of being in any of the pure states defined by the quantum walks evolutions.

The quantum Jensen-Shannon kernel $k_T(G_1, G_2)$ between the unattributed graphs G_1 and G_2 is defined as

$$k_T(G_1, G_2) = D_{JS}(\rho_T, \sigma_T) \tag{9}$$

where ρ_T and σ_T are the density operators defined as in Eq. 8. Note that this kernel is parametrised by the time T. In [1] the authors we propose to let $T \to \infty$, however, they show that a proper choice of T can yield an increased average accuracy in an SVM classification task.

It can be proved [1] that $0 \leq k_T(G_1, G_2) \leq 1$ and that if G_1 and G_2 are two isomorphic graphs, then ρ_T and σ_T have support on orthogonal subspaces, and as a consequence $k_T(G_1, G_2) = 1$. Note that although the authors are unable to provide a proof that the QJSD kernel is positive semidefinite, both empirical evidence and the fact that the Jensen-Shannon Divergence is negative semidefinite on pure quantum states [17] while the QJSD is maximal on orthogonal states suggest that it might be.

Fig. 2. Sample images of the four selected object from the COIL-100 [18] dataset

3.1 Enhancing the QJSD through Manifold Learning

Figure 1 shows the MDS embedding of the distance matrices associated with the QJSD kernel for the synthetic, MUTAG and COIL datasets. Details on the datasets used in this paper can be found in Section 4. These embeddings clearly suffer from a horseshoe shape effect, which is usually the result of an accurate estimate of the distance between objects only when they are close together, but not when they are far apart [16]. As a consequence, it should be possible to increase the kernel performance by filtering out in some way this long range distance information.

In this paper we propose a simple yet effective way to achieve this goal. Given a set of graphs, we compute the Isomap [14] embedding of the graphs and we evaluate the separability of the graph classes as the distance information varies from local to global. Isomap is a well-known manifold learning technique, which extends classical MDS by incorporating the pairwise geodesic distances between points. To this end, a neighborhood graph is constructed from the original set of points, where each node is connected to its k nearest neighbors in the high-dimensional space. The geodesic distance between two nodes is then defined as the sum of the edge weights along the shortest-path between them. It is known that Isomap suffers from several shortcomings, so further work should focus on experimenting with more robust manifold learning techniques.

The class separability is evaluated in the following way. For each embedding, we perform a 10-fold cross validation using a binary C-SVM with a linear kernel, where we let the value of the SVM regularizer constant C vary over the interval 10^{-3} and 10^{3}. Then, we take the maximum value of the average classification accuracy as an indicator of the separability. More formally, we look for the Isomap embedding which maximises

$$\arg\max_{d,k} \max_{C} \alpha \tag{10}$$

where α is the 10-fold cross validation accuracy of the C-SVM, C is the regularizer constant, d is the embedding dimension and k is the number of nearest neighbors. Note that the multi-classification task is solved using majority voting on a set of one-vs-one C-SVM classifiers.

4 Experimental Results

The experiments are performed on four different dataset, namely MUTAG, PPI, COIL [18] and a set of shock graphs. MUTAG is a dataset of 188 mutagenic

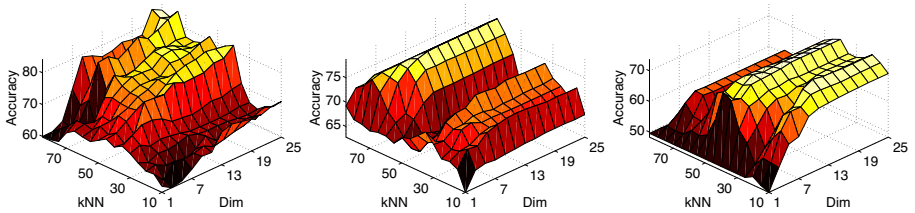

Fig. 3. 3D plot of the 10-fold cross validation accuracy on the PPI dataset as the number of the nearest neighbors k and the embedding dimension d vary

aromatic and heteroaromatic compounds labeled according to whether or not they have a mutagenic effect on the Gram-negative bacterium *Salmonella typhimurium*. The PPI dataset consists of protein-protein interaction (PPIs) networks related to histidine kinase from two different groups: 40 PPIs from *Acidovorax avenae* and 46 PPIs from *Acidobacteria*. The COIL dataset consists of the 4 objects shown in Figure 2, each with 72 views obtained from equally spaced viewing directions over 360°. For each image, a graph is obtained as the Delaunay triangulation of the Harris corner points. Finally, we select a set of shock graphs, a skeletal-based representation of the differential structure of the boundary of a 2D shape. The 120 graphs are divided into 8 classes of 15 shapes each. Each graph has a node attribute that reflects the size of the boundary feature generating the corresponding skeletal segment. To reflect the presence of attributes, the QJSD kernel is modified by labeling the new connections of the merged graph with the similarity between its two endpoints. To these four datasets, we add a fifth set of 30 synthetically generated graphs, 10 for each class. The graphs belonging to each class were sampled from a generative model with size 12,14 and 16 respectively [19].

Figure 3 shows the 3D plots of the 10-fold cross validation accuracy on the Isomap embeddings of the QJSD, the random walk and the graphlet kernels for the PPI dataset, as the size of the initial neighborhood and the embedding dimension vary. The plots show that for this dataset the QJSD kernel seems to be less sensitive to the locality of the distance information. On the other

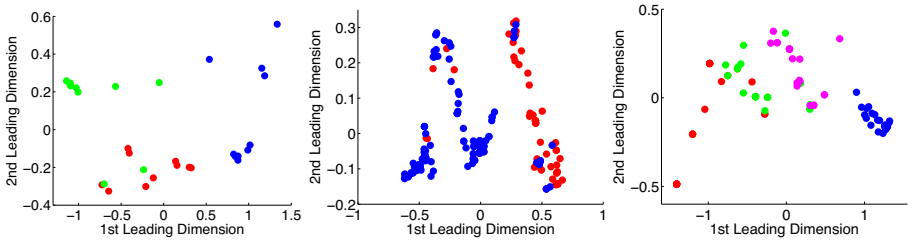

Fig. 4. The optimal two-dimensional Isomap embeddings in terms of separability between the graph classes

Table 1. Maximum classification accuracy on the unattributed graph datasets. Here SP is the shortest-path kernel of Borgwardt and Kriegel [6], RW is the random walk kernel of Gartner et al. [5], while GR denotes the graphlet kernel computed using all graphlets of size 3 described in Shervashidze et al. [7], while the subscript ISO indicates the result after the Isomap embedding. For each dataset, the best performing kernel before and after the embedding is shown in bold and italic, respectively.

Kernel	Synthetic	MUTAG	PPI	COIL	Shock
QJSD	*90.00*	*88.27*	*78.75*	84.44	*67.50*
QJSD_{ISO}	**96.67**	**91.96**	**90.69**	**91.53**	**77.50**
SP	80.00	86.08	71.25	85.56	61.67
SP_{ISO}	86.67	89.33	87.08	89.17	60.05
RW	86.67	77.02	70.97	79.72	49.17
RW_{ISO}	86.67	81.35	82.50	80.97	50.12
GR	86.67	82.92	49.56	*86.67*	39.17
GR_{ISO}	90.00	84.53	77.08	87.78	54.17

hand, for the graphlet kernel the maximum accuracy is achieved for a smaller neighborhood, which means that in this case the long range distance information is less accurate.

Figure 4 shows the two-dimensional Isomap embeddings with the highest linear separability for the QJSD kernels on the synthetic dataset, MUTAG and COIL. The result clearly shows the lack of the horseshoe shape distribution of Figure 1. Note, however, that the best embedding is usually found at a dimension higher than two and, as shown in Figure 3, the separability can change significantly as the dimension varies. Figure 4 also shows a clearer separation among the different classes, as highlighted in Table 1, which shows the separability of the data for each kernel and dataset. It is interesting to observe that, with the exception of a few cases, the Isomap embedding always yields an increased separability of the data, independently of the original kernel. It should also be underlined that the QJSD kernel always yields the highest separation, with a maximum classification accuracy above 90% in 4 out of 5 datasets.

5 Conclusions

In this paper, we studied the separability properties of the QJSD kernel and we have proposed a way to compute a low-dimensional embedding where the separation of the different classes is enhanced. The idea stems from the observation that the multidimensional scaling embeddings on this kernel show a strong horseshoe shape distribution, a pattern which is known to arise when long range distances are not estimated accurately. Here we proposed to use Isomap to embed the graphs using only local distance information onto a new vectorial space with a higher class separability. An extensive experimental evaluation has shown the effectiveness of the proposed approach.

Acknowledgments. Edwin Hancock was supported by a Royal Society Wolfson Research Merit Award.

References

1. Rossi, L., Torsello, A., Hancock, E.R.: A continuous-time quantum walk kernel for unattributed graphs. In: Kropatsch, W.G., Artner, N.M., Haxhimusa, Y., Jiang, X. (eds.) GbRPR 2013. LNCS, vol. 7877, pp. 101–110. Springer, Heidelberg (2013)
2. Siddiqi, K., Shokoufandeh, A., Dickinson, S., Zucker, S.: Shock graphs and shape matching. International Journal of Computer Vision 35, 13–32 (1999)
3. Jeong, H., Tombor, B., Albert, R., Oltvai, Z., Barabási, A.: The large-scale organization of metabolic networks. Nature 407, 651–654 (2000)
4. Schölkopf, B., Smola, A.J.: Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT press (2001)
5. Gaertner, T., Flach, P., Wrobel, S.: On graph kernels: Hardness results and efficient alternatives. In: Schölkopf, B., Warmuth, M.K. (eds.) COLT/Kernel 2003. LNCS (LNAI), vol. 2777, pp. 129–143. Springer, Heidelberg (2003)
6. Borgwardt, K., Kriegel, H.: Shortest-path kernels on graphs. In: Fifth IEEE International Conference on Data Mining, p. 8. IEEE (2005)
7. Shervashidze, N., Vishwanathan, S., Petri, T., Mehlhorn, K., Borgwardt, K.: Efficient graphlet kernels for large graph comparison. In: Proceedings of the International Workshop on Artificial Intelligence and Statistics (2009)
8. Haussler, D.: Convolution kernels on discrete structures. Technical report, UC Santa Cruz (1999)
9. Farhi, E., Gutmann, S.: Quantum computation and decision trees. Physical Review A 58, 915 (1998)
10. Emms, D., Wilson, R., Hancock, E.: Graph embedding using a quasi-quantum analogue of the hitting times of continuous time quantum walks. Quantum Information & Computation 9, 231–254 (2009)
11. Rossi, L., Torsello, A., Hancock, E.R.: Approximate axial symmetries from continuous time quantum walks. In: Gimel'farb, G., Hancock, E., Imiya, A., Kuijper, A., Kudo, M., Omachi, S., Windeatt, T., Yamada, K. (eds.) SSPR&SPR 2012. LNCS, vol. 7626, pp. 144–152. Springer, Heidelberg (2012)
12. Lamberti, P., Majtey, A., Borras, A., Casas, M., Plastino, A.: Metric character of the quantum Jensen-Shannon divergence. Physical Review A 77, 052311 (2008)
13. Nielsen, M., Chuang, I.: Quantum computation and quantum information. Cambridge university press (2010)
14. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323 (2000)
15. Czaja, W., Ehler, M.: Schroedinger eigenmaps for the analysis of biomedical data. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 1274–1280 (2013)
16. Kendall, D.G.: Abundance matrices and seriation in archaeology. Probability Theory and Related Fields 17, 104–112 (1971)
17. Briët, J., Harremoës, P.: Properties of classical and quantum jensen-shannon divergence. Physical review A 79, 052311 (2009)
18. Nayar, S., Nene, S., Murase, H.: Columbia object image library (coil 100). Technical report, Tech. Report No. CUCS-006-96. Department of Comp. Science, Columbia University (1996)
19. Torsello, A., Rossi, L.: Supervised learning of graph structure. In: Pelillo, M., Hancock, E.R. (eds.) SIMBAD 2011. LNCS, vol. 7005, pp. 117–132. Springer, Heidelberg (2011)