Distance based Kernels for First-Order Logic Data

Nirattaya Khamsemanan[1], Cholwich Nattee[1], and Masayuki Numao[2]

[1]Sirindhorn International Institute of Technology, Thammasat University, Thailand
{nirattaya,cholwich}@siit.tu.ac.th
[2]The Institute of Scientific and Industrial Research, Osaka University, Japan
numao@ai.sanken.osaka-u.ac.jp

Abstract. Support Vector Machines (SVM) and kernel techniques have been proven effective on various application domains using attribute-value representation. A number of works have been done to apply SVM on First-Order Logic (FOL) data as well as using SVM with Inductive Logic Programming (ILP). In this paper, we propose kernel functions for FOL data developed from the four-layer distance metric. Since our proposed kernels are not positive definite, we apply the shift spectrum transformation to ensure that the kernel matrices are positive semidefinite before use them in the SVM optimization algorithm. The proposed kernels yields higher accuracies than the baseline ILP system on Mutagenesis and Alzheimer dataset. They significantly outperform the existing kernel functions on the Alzheimer dataset. On the Mutagenesis dataset, our kernel functions performs not significantly different from the best accuracy.

1 Introduction

The real-world data cannot always be represented numerically. Many types of data are represented First-Order Logic (FOL) representation especially the structured data with relationships among objects. Various techniques have been invented to learn from this type of data. Inductive Logic Programming (ILP) is among those techniques. Support Vector Machines (SVM) have been used in the ILP community. The basic idea of SVM is to construct a maximal margin linear classifier. Kernel functions are necessary for SVM to deal with non-linearly separable datasets. These functions embed datasets to Hilbert spaces. To ensure the maximal margin classifiers, the kernel functions must be positive definite, or equivalently the kernal matrices must be positive semidefinite.

In Section 2, we describe the format of FOL datasets that are appropriate for our proposed kernel functions. Section 3 provides the definition of the four-layer distance metric. Section 4 shows how to construct four-layer distance based kernel functions. The results of the experiments are shown in Section 5. Finally, we conclude our work in Section 6.

2 Setting

We are working under the assumption that elements in a dataset is defined as follow:

Definition 1. *A (first-order logic) dataset $\mathcal{C}$ is a set of elements of the form*

$$ID^r = r(ID, x_1, ..., x_n),$$

where r is a predicate symbol. The first entry of an atom will be called rank 0, the $(i+1)^{th}$ entry will be call rank i of an atom. An argument in the rank 0 is called ID. An element of $\mathcal{C}$ is called an atom.

Definition 2. *An object X is an FOL object if its entire properties and identities are can be expressed by atoms in a (first-order logic) dataset $\mathcal{C}$. A set of all atoms with $ID = X$ is called the property set of X.*

A dataset $\mathcal{C}$ is then a union of property sets of FOL objects. A single-level structure object is a FOL object whose ID argument will not appear in rank 0 of other atoms. A multi-level structure object is a FOL object whose arguments in all entries of any atoms can be ID of other atoms.

3 The Four-Layer Distance Metric

Definition 3. *Suppose X and Y are two FOL objects whose properties are represented in a dataset $\mathcal{C}$ where both X and Y are main IDs. The four layer distance between X and Y is defined as follow:*

Layer 1: ***Distance function of arguments with respect to a predicate symbol r and rank i:*** *Suppose $X^r = r(X, x_1, x_2, \cdots, x_n)$ and $Y^r = r(Y, y_1, y_2, \cdots, y_n)$ are two atoms in $\mathcal{C}$ with the same predicate symbol r. The distance between x_i and y_i is defined as*

$$\Delta_{r,i}(x_i, y_i) = \begin{cases} 0 & \textit{if } x_i = y_i, \\ 1 & \textit{if } x_i \neq y_i, \textit{ and } x_i \notin \mathbb{R} \textit{ or } y_i \notin \mathbb{R}, \\ \dfrac{|x_i - y_i|}{\max(r,i)} & \textit{if } x_i \neq y_i \textit{ and } x_i, y_i \in \mathbb{R}. \end{cases}$$

where $\max(r,i)$ is the maximum difference of all pairs of arguments in the rank i of atoms with the predicate symbol r, ranging over $\mathcal{C}$.

Layer 2: ***Distance function of atoms with respect to a predicate symbol r:*** *Suppose $X^r = r(X, x_1, x_2, \cdots, x_n)$ and $Y^r = r(Y, y_1, y_2, \cdots, y_n)$ are two atoms in $\mathcal{C}$ with the same predicate symbol r. The distance function of two atoms with the same predicate symbol r is defined as*

$$d_r(X^r, Y^r) = \sqrt{\frac{\sum_{i=1}^{n} \left(\delta_{r,i}(x_i, y_i)\right)^2}{n}}$$

where,

$$\delta_{r,i}(x_i, y_i) = \begin{cases} 0 & \textit{if } x_i = y_i, \\ \Delta_{r,i}(x_i, y_i) & \textit{if at most one of } x_i, y_i \textit{ is an ID}, \\ D(x_i, y_i) & \textit{if both } x_i, y_i \textit{ are ID's}. \end{cases}$$

Note that in the case where x_i and y_i are not equal and are both IDs of other atoms, then $\delta_{r,i}(x_i, y_i)$ is set to $D(x_i, y_i)$, ranging over Ω', a set of all predicate symbols of atoms whose IDs are arguments in rank i of any atoms with the predicate symbol r and is defined recursively until x_i, y_i are single-level structure objects.

Layer 3: ***Distance between two FOL objects with respect to a predicate symbol*** r***:*** *For a predicate symbol r, suppose there are p atoms with main $ID = X$ and q atoms with main $ID = Y$, then the r-distance between an object X and an object Y is*

$$D_r(X,Y) = \begin{cases} \max\{\max\limits_{k=1}^{p}\min\limits_{j=1}^{q} d_r(X^{r_k}, Y^{r_j}), \max\limits_{j=1}^{q}\min\limits_{k=1}^{p} d_r(X^{r_k}, Y^{r_j})\} & \text{\textit{if} } p, q \neq 0 \\ 1 & \text{\textit{if} } p \neq 0, q = 0, \\ & \quad \text{\textit{or}} \;\; p = 0, q \neq 0 \\ 0 & \text{\textit{if} } p = q = 0 \end{cases}$$

Layer 4: ***The Four-layer Distance between two FOL objects:*** *For two objects main $ID = X$ and main $ID = Y$ whose properties are expressed as atoms in $\mathcal{C}$. The distance between X and Y is defined as*

$$D(X,Y) = \sqrt{\frac{\sum\limits_{r \in \Omega} (D_r(X,Y))^2}{|\Omega|}}$$

where Ω is the set of predicate symbols of atoms that contain main ID in $\mathcal{C}$. In the calculation of $D(x_i, y_i)$, if at some level, $D(x_i, y_i)$ is a function of itself, then $D(x_i, y_i)$ is obtained by solving that equation. This equation where $D(x_i, y_i)$ is a function of itself is call the distance equation of $D(x_i, y_i)$.

Note that the four-layer distance function in this section is defined recursively. This will allow our four-layer distance function to support multi-level structure datasets.

Theorem 1. *The four-layer distance function is a metric.*

4 Four-Layer Distance Based Kernels

A kernel $k : \mathcal{C} \times \mathcal{C} \to \mathbb{R}$ is a real-valued function that takes a cartesian product of elements in $\mathcal{C}$ and returns a real number. If k is positive definite, then there exist a map Φ that isometically embeds $\mathcal{C}$ into a Hilbert space $\mathcal{H}$ such that,

$$\langle \Phi(X), \Phi(Y) \rangle = k(X,Y), \tag{1}$$

which is a dot product of $\Phi(X)$ and $\Phi(Y)$ in $\mathcal{H}$. Because of the equation 1, the computation of dots product in $\mathcal{H}$ can be done on $\mathcal{C}$ without needing to know what $\Phi(X)$ and $\Phi(Y)$ are. In SVM, the positive definite property of k is required to secure the maximal margin in the Hilbert space $\mathcal{H}$.

A distance based kernel is a kernel that are created based on a distance metric d such that $d(X,Y) = ||\Phi(X) - \Phi(Y)||_{\mathcal{H}}$.

Let $D(X,Y)$ be a four-layer distance metric defined in Definition 3. We define a kernel function as follow:

$$k(X,Y) = k_O(X,Y) = \frac{1}{2}\left(D(X,Y)^2 - D(X,O)^2 - D(Y,O)^2\right),$$

where O is a fixed object in a dataset $\mathcal{C}$.

Notice that,

$$\begin{aligned}
||\Phi(X)-\Phi(Y)||^2_{\mathcal{H}} &= \langle\Phi(X)-\Phi(Y)\rangle\langle\Phi(X)-\Phi(Y)\rangle \\
&= \langle\Phi(X),\Phi(X)\rangle - 2\langle\Phi(X),\Phi(Y)\rangle + \langle\Phi(Y),\Phi(Y)\rangle \\
&= \left(\frac{1}{2}\left(D(X,X)^2 - D(X,O)^2 - D(X,O)^2\right)\right) \\
&\quad -2\left(\frac{1}{2}\left(D(X,Y)^2 - D(X,O)^2 - D(Y,O)^2\right)\right) \\
&\quad \left(\frac{1}{2}\left(D(Y,Y)^2 - D(Y,O)^2 - D(Y,O)^2\right)\right) \\
&= D(X,Y)^2
\end{aligned}$$

We construct four different types of kernel function based on the four-layer distance metric (4L-kernels) [2]:

1. A four-layer distance based simple linear kernel:

$$k^{lin}_{4L}(X,Y) = \frac{1}{2}\left(D(X,Y)^2 - D(X,O)^2 - D(Y,O)^2\right),$$

where O is a fixed object in a dataset $\mathcal{C}$.

2. A four-layer distance based negative-distance kernel:

$$k^{nd}_{4L}(X,Y) = -\left(D(X,Y)^2\right).$$

3. A four-layer distance based polynomial kernel:

$$k^{pol}_{4L}(X,Y) = \left(1+\gamma\left(D(X,Y)^2 - D(X,O)^2 - D(Y,O)^2\right)\right)^p,$$

where $\gamma \in \mathbb{R}^+$ and $p \in \mathbb{N}$.

4. A four-layer distance based Gaussian kernel:

$$k^{gs}_{4L}(X,Y) = e^{-\gamma D(X,Y)^2},$$

where $\gamma \in \mathbb{R}^+$.

The kernel k is be positive definite if and only if the Gram matrix (kernel matrix) $K = [k(X^i, X^j)]$ where $i,j = 1,\ldots,|\mathcal{C}|$ must be positive semidefinte, i.e., eigenvalues of K are non-negative.

The 4L-kernels are, by themselves, not positive definite function because the 4L distance metric is not conditionally positive definite on some datasets. In the indefinite cases, we apply the *shift spectrum transformation* [6] on the indefinite Gram matrix to obtain the positive semidefinite one. The shift spectrum transformation is as follow:

$$\widetilde{K} = U\widetilde{\Lambda}U^T = U(\Lambda + \eta I)U^T = K + \eta I$$

Wu et al. [6], have shown that if η is greater than $|\lambda_N|$, where $\lambda \geq \lambda_N$ for all eigenvalues λ of K, then $\widetilde{K}$ is positive semidefinite. The shift spectrum transformation preserve semantic of the Gram matrix since it does not change the off-diagonal entries. They also showed that minimizing the dual formation after shift is equivalent to minimizing both the dual formation before shift and 2-norm of the multiplier vector.

5 Experiments and Discussions

We test our 4L-kernels on real-world ILP datasets, i.e. Mutagenesis dataset [5], and the Alzheimer dataset [3] with 4 different properties. The experiments are conducted by training Support Vector Machines (SVM) using 10-fold cross validation method. In each dataset, if the Gram matrix is indefinite, we apply the shift spectrum transformation with $\eta = |\lambda_N|$, where λ_N is the minimum eigenvalue as defined in the previous section. From the preliminary test, we found that the accuracy was achieved with $\gamma = 1$ in the Gaussian kernel and $\gamma = 1, p = 5$ for polynomial kernel. Thus, we use $\gamma = 1, p = 5$ to conduct the experiment.

We compare our results (k_{4L}^{lin}, k_{4L}^{nd}, k_{4L}^{pol}, k_{4L}^{gs}) with 4 types of kernels based on RB distance metric [4] (k_{RB}^{lin}, k_{RB}^{nd}, k_{RB}^{pol}, k_{RB}^{gs}), and the kernel for structured data [1] (k_{DK}). The results from Aleph[1], one of the most widely used ILP systems, is used as the baseline. In RB distance based kernels, we use the same parameter $\gamma = 1, p = 5$ in Gaussian and polynomial kernels as suggested by a preliminary experiment. The results of the experiment are shown in Table 1. Since k_{DK} has been proven to be a positive definite kernel, we do not apply the spectrum transformation on the kernel matrices from this kernel.

The results show that SVM using our proposed 4L-kernels perform better than Aleph, the baseline technique for ILP. Our kernel k_{4L}^{pol} yields the best accuracy on all four properties of the Alzheimer dataset. The accuracies are significantly different from a number of methods. The kernel k_{RB}^{gs} yields the best accuracy on the Mutagenesis dataset. However k_{RB}^{gs} is not significantly difference from our kernel k_{4L}^{pol} in this dataset. The results also show that all types of 4L-kernels perform generally better than the other techniques.

Kernel matrices from our proposed 4L-kernels on the *amine* property of the Alzheimer dataset are positive semidefinite. They perform significantly better than the other indefinite kernels.

6 Conclusion

We propose the kernel functions of FOL data based on the four-layer distance metric. Since the kernel functions are not positive definite on some datasets, we apply the shift spectrum transformation technique to make the kernel matrices be positive semidefinite. We evaluate the proposed kernel functions by using

[1] `http://www.cs.ox.ac.uk/activities/machlearn/Aleph/aleph_toc.html`

Table 1. Prediction accuracies on real-world datasets using 10-fold cross validation method (figures in boldface font indicate the best accuracy in each dataset, the symbol † indicates the positive semidefinite kernel matrix, and the accuracies with the symbol * are significantly different from the best accuracy in their dataset with $p < 0.01$.

Method	Muta	Alz amine	Alz toxic	Alz acetyl	Alz memory
Aleph	73.4 ± 11.8	$70.2 \pm 7.3^*$	$90.9 \pm 3.5^*$	$73.5 \pm 4.3^*$	$69.3 \pm 3.9^*$
k_{4L}^{lin}	70.2 ± 12.4	$92.1 \pm 4.5^*$	$94.4 \pm 2.1^*$	$93.0 \pm 1.6^*$	87.4 ± 5.6
k_{4L}^{nd}	72.8 ± 8.1	$93.3 \pm 5.3^*$	98.0 ± 1.6	$93.0 \pm 2.9^*$	88.5 ± 3.8
k_{4L}^{pol}	79.7 ± 8.1	$\mathbf{96.4 \pm 2.7}^\dagger$	$\mathbf{98.2 \pm 1.1}$	$\mathbf{95.6 \pm 2.0}$	$\mathbf{89.6 \pm 3.9}$
k_{4L}^{gs}	74.0 ± 9.3	$93.0 \pm 4.6^{\dagger *}$	$95.5 \pm 2.2^*$	$92.8 \pm 2.0^*$	88.3 ± 4.8
k_{RB}^{lin}	82.4 ± 6.3	$72.6 \pm 4.3^*$	$62.7 \pm 4.4^*$	$62.7 \pm 3.0^*$	$52.5 \pm 8.4^*$
k_{RB}^{nd}	81.9 ± 8.3	$70.8 \pm 5.7^*$	$59.9 \pm 5.3^*$	$62.0 \pm 3.8^*$	$52.2 \pm 8.3^*$
k_{RB}^{pol}	77.6 ± 5.2	$72.5 \pm 4.8^*$	$73.3 \pm 5.1^*$	$69.8 \pm 4.1^*$	$59.5 \pm 3.6^*$
k_{RB}^{gs}	$\mathbf{83.6 \pm 9.7}$	$85.1 \pm 3.8^*$	$88.5 \pm 2.6^*$	$81.3 \pm 3.0^*$	$74.3 \pm 3.0^*$
k_{DK}	82.0 ± 10.8	$93.2 \pm 3.5^*$	$96.4 \pm 1.8^*$	94.6 ± 2.3	88.8 ± 3.7

them to construct SVM for real-world datasets. The experimental results show that SVM with the proposed kernel functions provide higher accuracy than the baseline ILP system, and generally perform better than existing kernel functions.

7 Acknowledgement

This work was performed under the Cooperative Research Program of "Network Joint Research Center for Materials and Devices" at the Institute of Scientific and Industrial Research, Osaka University, Japan. It was also supported by the Institute for the Promotion of Teaching Science and Technology (IPST) and Thammasat University Research Fund, Thailand.

References

1. Gärtner, T., Lloyd, J.W., Flach, P.A.: Kernels and distances for structured data. Machine Learning 57, 205–232 (2004)
2. Haasdonk, B., Bahlmann, C.: Learning with distance substitution kernels. In: Rasmussen, C., BÃijlthoff, H., SchÃűlkopf, B., Giese, M. (eds.) Pattern Recognition, Lecture Notes in Computer Science, vol. 3175, pp. 220–227. Springer Berlin Heidelberg (2004), `http://dx.doi.org/10.1007/978-3-540-28649-3_27`
3. King, R., Srinivasan, A., Sternberg, M.: Relating chemical activity to structure: an examination of ilp successes. New Generation Computing 13(2, 4), 411–433
4. Ramon, J., Bruynooghe, M.: A polynomial time computable metric between point sets. Acta Informatica 37(10), 765–780 (2001)
5. Srinivasan, A., Muggleton, S., King, R., Sternberg, M.: Theories for mutagenicity: a study of first-order and feature-based induction. Artificial Intelligence 85, 277–299
6. Wu, G., Chang, E.Y., Zhang, Z.: An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines. In: Proceedings of the 22nd International Conference on Machine Learning (2005)