

A NETFLOW DISTANCE BETWEEN LABELED GRAPHS: APPLICATIONS IN CHEMOINFORMATICS

Subhransu Maji

Department of Computer Science,
Indian Institute of Technology, Kanpur
email: maji@cse.iitk.ac.in

Shashank Mehta

Department of Computer Science,
Indian Institute of Technology, Kanpur
email: skmehta@cse.iitk.ac.in

ABSTRACT

We propose a novel measure of similarity between labeled graphs which has applications to structured data analysis, for e.g. chemical informatics, web document clustering, etc. Exact metrics on graphs based on sub-graph isomorphism have been proposed earlier but due to the lack of an efficient algorithm, they cannot be applied on large sized data. Our metric on graphs exploits vertex context similarity and computes an overall matching score in polynomial time in the size of the graphs using a network flow formulation of the problem. We use this metric, in a discriminative framework for predicting chemical properties like carcinogenicity and mutagenicity of molecules and test it on PTC and MUTAG datasets. The positive definite kernels constructed using this metric shows significantly improved performance over existing kernels for graphs on most datasets demonstrating the efficacy of the technique.

KEY WORDS

Classification, Support Vector Machines, Graph Kernels, Drug Activity Prediction.

1 Introduction

Many of the real world data is structured and have a natural representation as graphs. These include problems of information retrieval from text documents, DNA/RNA/protein sequences and evolutionary trees in bioinformatics as well as molecular structures in chemical informatics. In this work we develop a technique to classify chemical data represented in the form of graphs with labeled vertices and edges based on graph based similarity measures. A technique for predicting the activity of drug molecules has the potential of saving the huge costs and time incurred to conduct the actual experiments in the laboratory.

The metric we choose to measure the similarity between the graphs must be chemically motivated. Chemical properties are captured well by locally describable properties like atom labels, bond types and functional groups. To test this hypothesis we conduct a small experiment. We rank various chemical features of a compound for the task of chemical property classification. Some of these features are local properties,

i.e. its presence can be determined by looking at a small portion of the chemical molecule, while the rest are global properties, i.e. is the property of the entire molecule.

We use the Predictive Toxicology Challenge (PTC) dataset [1] is used which reports the reports the carcinogenicity of several hundred chemical compounds for Male Mice (MM)(The full dataset is used later for the evaluating the classifier performance). The features include the atom counts of various atoms in the molecule (21 different atom types) as well as the counts of the following properties; *Aromatic, Ring, Alkane, Alkene, Alkyne, Amine, Carbonyl, Ester, CarboxylicAcid, Anhydrid, Amide, AlkylHalide, Halide, Aldehyde, Ketone, Alcohol, Thiol, Ether, ThioEther, Fenol, AmineSalt, Imine, Nitrile, IsoCyanate, Nitro, Acetale, SulfonicAcid, SulfonicAmide, PhosphateEster, PhosphorAmidate* obtained from <http://www.predictive-toxicology.org/>. The features were ranked using the SVM Attribute Evaluator in WEKA[2]. The results are described in the Figure 1. We also compare the performance of the various classifiers which use the high level knowledge of the presence or absence of the above properties, with other popular techniques. The accuracies of various learning algorithms over these features using a 10 fold cross validation are shown in Figure 2.

We notice that most of the highly ranked feats are local features. Also, the accuracies obtained are better than the existing kernels for graphs. This motivates us to have a two fold approach for matching in which we compare the graphs first at a local scale followed by a holistic matching. A distance measure ideally should have the properties of a metric, while also should be easy to compute. The rest of the paper is structured as follows. We outline the previous work in section 2, in section 3 we introduce our graph metric, in section 4 and 5, we outline our admissible graph kernel and present the experiments and results on two different datasets.

2 Previous Work

Measuring the similarity or distance between graphs as we have seen is an important problem in the ma-

Rank	Attribute	Diameter	Type
1	sn	1	local
2	IsoCyanate	3	local
3	ba	1	local
4	f	1	local
5	Halide	1	local
6	br	1	local
7	cl	1	local
8	i	1	local
9	h	1	local
10	c	1	local
11	Aromatic	-	global
12	Alkane	-	global
13	Alkene	2	local
14	Ether	3	local
15	s	1	local
16	Nitrile	2	local
17	Carbonyl	2	local
18	o	1	local
19	Nitro	3	local
20	Alcohol	2	local

Figure 1. The top 20 ranked features are shown in the table. Diameter denotes the diameter of the context around the center atom which encloses the corresponding property. Notice that most of the properties are local. Aromaticity is property arising out of conjugations, etc, while alkaneness is property which can be found only by looking into the entire molecule.

Learner	Accuracy
SMO (SVM with a linear Kernel)	69.74%
Voted Perceptron	71.06 %
J4.8	67.99%
Marginalized	64.3%
Tanimoto	66.4%
MiniMax	64.0%

Figure 2. The first three rows are the accuracies using the corresponding learning algorithms. The last three are the accuracies of three best known graphs kernels in literature.

chine learning community. Measures based on Edit Distance have been proposed by [3]. Here one introduces a set of edit operations like insertion, deletion and substitution of nodes and edges and the distance between the graphs is the minimum number of these edit operations required to transform one graph to the other. Similarity based on the maximal common subgraphs have been suggested in [4, 5] and go on to show that the distance is a metric. Classical methods for computing the maximal common subgraph are based on maximal clique detection. Though conceptually simple these methods have high complexity hence are prohibitive for graphs of large size. For chemical compounds similarities based on various representations including fingerprints(2D and 3D)[6], connection tables and physiochemical property vectors(“dataprints”) are well known.

In recent years, kernel methods have emerged as an important class of machine learning methods suitable for variable-size structured data [7]. Given two input objects u and v , such as two molecules, the basic idea behind kernel methods is, to construct a kernel $k(u, v)$ which measures the similarity between u and v . This kernel can also be viewed as an inner product of the form $k(u, v) = \langle \phi(u), \phi(v) \rangle$ in an embedding feature space determined by the map, which need not be given explicitly.

Kernels that can be applied to graphs have recently been introduced in and have been successfully applied to the task of biomolecular activity prediction. A class of Kernels based on the powers of adjacency Matrix was proposed by [8], while *labeled-pair kernels* count the number of walks in both the graphs of the same length and with the same labels on their first and last nodes. Intuitively ”similar” graphs should share large number of such labeled walks than two ”dissimilar” ones. A variant of this is the class of kernels based count the number of labeled sequences both the graphs share [9].

Marginalized Graph Kernels [10] compute the similarity of two graphs from the probability of two random walks being same on each of the graphs. The kernel is parameterized by transition probabilities between nodes and the probability of a walk halting at every node. Further improvements on the marginalized kernels have been done in [11] which incorporates local neighborhood information and removes totters. In spite of all these computing these Marginalized Kernels requires an iterative procedure which is computationally expensive. Recently two new kernels called the Tanimoto and MiniMax[6] kernels have been proposed and their performance was better on most cases than the previous kernels on the task of biomolecular classification.

3 Metric on Graphs

Based on the the motivation of local matching we propose distance between labeled graphs as follows: a graph, G , is broken up into small overlapping subgraphs, each representing a neighboring region or the context around a vertex. Thus, each graph is represented by a set of subgraphs. We take a two step approach for the similarity computation; first define a distance between the subgraphs, and second define a overall distance between two sets of subgraphs. We discuss the choices for each of these in the following subsections.

3.1 Metric on Subgraphs

Various possibilities including the ones described earlier can be used. In our experiments we use the following two measures:

Maximal Common Subgraph Based Measures

These measures proceed by computing the maximal common subgraph between the graphs and then defining the distance as

$$d(G_1, G_2) = 1 - \frac{|G_{12}|}{\max(|G_1|, |G_2|)} \quad (1)$$

where, $|G_{12}|$, $|G_1|$ and $|G_2|$ are the sizes of common subgraph and graphs G_1 and G_2 respectively. This distance has been shown to be a metric in [4]. Two notions of sizes can be used namely the maximal common subgraph and the maximal common edge induced graphs. Computing each of these is tractable as long as each of the subgraphs are small in size.

Atom Label Histograms Based Measures

This first transforms every subgraph to a multidimensional vector corresponding to the frequencies of the atom labels in it and then computing the distance between these vectors. For our experiments, we have used the L1 Normalized distance $d^n(\vec{X}, \vec{Y})$ as the metric

$$d^n(\vec{X}, \vec{Y}) = \sum_{i=1}^N |X_i - Y_i| \quad (2)$$

where the vectors are first L1 normalized according to the equation $\vec{V} \rightarrow \vec{V}/(|\vec{V}|_1 + \epsilon)$. Computing this distance is linear in the size of the subgraphs.

3.2 Metric on subgraph sets

Once we have metric between the subgraphs we need to extend this to a metric between two sets of subgraphs. Several choices exist in literature, including the Hausdorff metric. This, as observed by many does not take

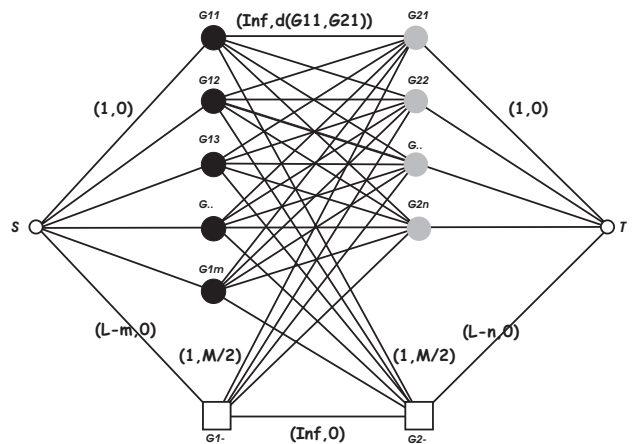


Figure 3. The flow network corresponding to the matching problem. The labels on the edges are (capacity, weight) tuples. L is taken to be greater than the size of the largest graph in the dataset, while M (penalty term for no match) is taken to the largest distance between the subgraphs.

into account the structure of the point sets (determined by the most distant element of both sets to the nearest neighbour in the other set). This makes the metric unsuited for applications where one set is likely to contain a point which is very different from other points in the set. To take into account the structure propose to formulate this problem as a bipartite matching problem. Given two graphs \mathbf{G}_1 and \mathbf{G}_2 , let,

$$\mathbf{G}_1 =_d \{G_{11}, \dots, G_{1n}\} \text{ and } \mathbf{G}_2 =_d \{G_{21}, \dots, G_{2m}\} \quad (3)$$

denote the decomposition of the graphs into subgraphs. Here n and m are the number of vertices of \mathbf{G}_1 and \mathbf{G}_2 respectively. We can define the distance between the sets as the weight of the optimal matching between the sets. We can formulate this as the matching defined by the distance network as shown in figure 3. This distance is shown to be a metric [12]. The metric is equal to the weight of a maximal flow of min weight in this flow network, which can be computed efficiently.

4 Classification using Kernel SVMs

Here we deal with the tools we require to use the metric constructed to build a classifier for our task. We train a *Support Vector Machine* (SVM) for the purpose of molecular classification. The basic theory of SVMs can be found in [13]. SVMs perform pattern recognition for a two class problem by determining the separating hyperplane with the maximum distance to the closest point of the training set. These points are called support vectors. If the data is not linearly separable in the input space, a non-linear transforma-

tion can be applied which maps the data points in the input space into a high(possibly infinite) dimensional space which is called feature space. The data in the feature space is then separated by the optimal hyperplane as described above. An attraction of using the SVMs is that instead of an explicit representation of data as vectors, it only requires inner products between such representations, through what is usually referred to as a positive definite kernel function. The condition of positive definiteness ensures that there exists a mapping of the data to some hilbert space of features(of possibly infinite dimension) for which the kernel represents the dot product in that feature space, i.e. $k(u, v) = \langle \phi(u), \phi(v) \rangle$.

4.1 A Positive Definite Kernel based on the Graph Metric

We construct a Positive Definite Kernel using the metric we discussed as follows. The features are constructed based on the distances to other graphs, i.e.,

$$\phi(x) = (d(x_1, x), \dots, d(x_n, x))$$

This feature is a vector in \mathfrak{R}^n , and we use the RBF kernel which known to be positive definite over the features. Let $x, x' \in \mathfrak{R}^n$ and let $\langle x, x' \rangle$ denote the scalar product in \mathfrak{R}^n . Given the bandwidth parameter σ the RBF Kernel is defined as:

$$k(x, x') = e^{-\frac{\|x-x'\|^2}{\sigma^2}} \quad (4)$$

We use this kernel derived from the graph distance metric based on optimal matchings for learning the classification of graphs.

5 Data Collection and model testing Drug Activity Prediction

A large number of datasets are available for the task of classification of drug molecules. Drug molecules are naturally represented a graphs with labeled vertices and edges. We have tested the kernels developed in the earlier sections for the task of predicting toxicity and mutagenicity on two different datasets(PTC and Mutag). Both these datasets are popular in literature [1, 14], so we also use it to provide a comparison with other existing methods for the same task. The details of the datasets are as follows.

PTC Dataset

The Predictive Toxicology Challenge (PTC) dataset [1] reports the carcinogenicity of several hundred chemical compounds for Male Mice (MM), Female Mice (FM), Male Rats (MR) and Female Rats (FR).

According to their carcinogenicity, each of the compounds are labeled one of the following labels: {EE, IS, E, CE, SE, P, NE, N} where CE, SE and P are "relatively active", and NE and N are "relatively inactive", and EE, IS and E indicate "cannot be decided". In order to simplify our problem we label CE, SE and P as positive while NE and N are taken to be negative. The task is to predict whether each molecule is positive or not for each of the four animal species.

Mutag Dataset

The Mutag dataset [14] consists originally of 230 chemical compounds assayed for mutagenicity in Salmonella typhimurium (Table 1). Among the 230 compounds, however, only 188 (125 positive, 63 negative) are considered to be learnable [14] and thus are used in the simulations. Various statistics for these datasets are summarized in Table 1.

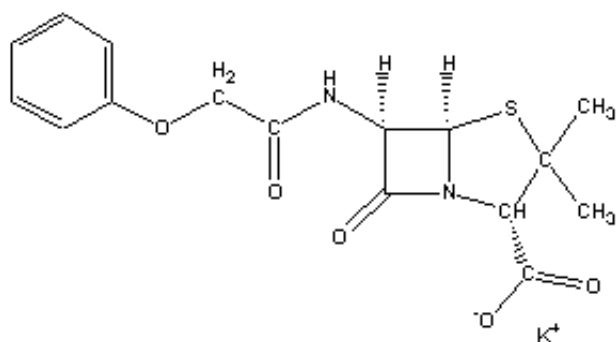
Feature	MM	FM	MR	FR	Mutag
# positive	129	143	152	121	125
# negative	207	206	192	230	63
max. $ G $	109	109	109	109	40
avg. $ G $	25.0	25.2	26.1	26.1	31.4
max. degree	4	4	4	4	4
$ \Sigma _v$	21	19	19	20	8
$ \Sigma _e$	4	4	4	4	4

Table 1. Several statistics of the dataset like the number of positive examples(#positive), number of negative examples (#negative), maximum degree(max. degree), maximum size of graphs(max. $|G|$), average size of the graphs (avg. $|G|$), number of vertex labels ($|\Sigma|_v$) and number of vertex labels ($|\Sigma|_e$).

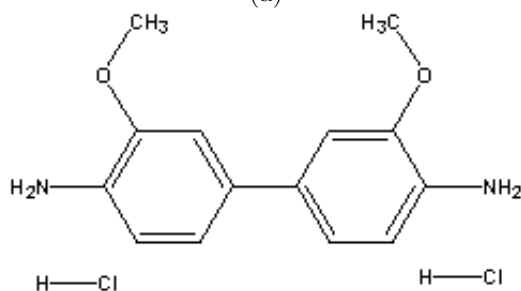
5.1 Experimental results

We train the classifier using both the notions of distance based on maximal edge induced and vertex induced subgraphs. The context radius (r), is taken to be 1. The classification accuracies are reported in Table 2 by leave one out cross validation. The accuracies(in %) for different choices of the similarity measures are summarized in the table. The last three rows are the results obtained by using the marginalized kernel[10], Tanimoto and Minimax Kernel[6], who also have used the leave one out cross validation technique.

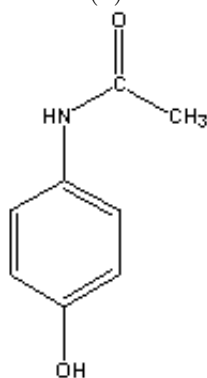
Table 3 summarizes the results using the histogram kernels. As the computation of this kernel is considerably less complex than the subgraph based ones, we evaluate the performance for higher values of context radius 1, 2, 3 and 4. The diameter of the molecular graphs most molecular graphs do not exceed 10 so higher values of the radius are not considered($r = 5$ would correspond to the entire graph). There is not



(a)



(b)



(c)

Figure 4. Sample molecules from the PTC dataset. (a) $C_{16}H_{17}KN_2O_5S$ Penicillin VK, (b) $C_{14}H_{16}N_2O_2 \cdot 2HCl$ 3,3'-Dimethoxybenzidine Dihydrochloride, (c) $C_8H_9NO_2$ Acetaminophen. The entire collection of molecules can be found at <http://www.predictive-toxicology.org/ptc/>

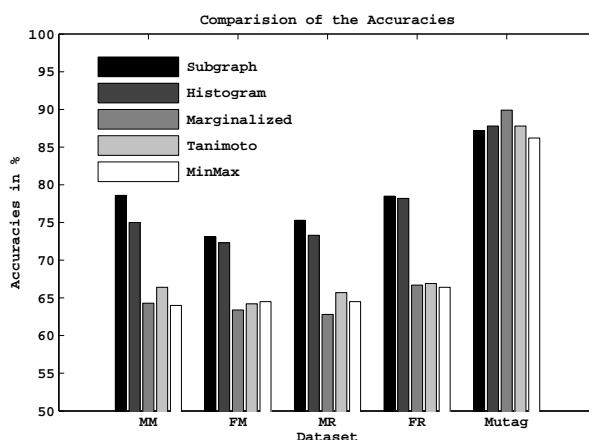


Figure 5. A bar plot of accuracies in % of various techniques on different datasets. The best of each class for the subgraph and histograms are taken as representatives.

kernel	MM	FM	MR	FR	Mutag
<i>MCS</i> ($r = 1$)	76.8	72.4	75.3	78.5	84.1
<i>MCES</i> ($r = 1$)	76.5	73.1	72.9	78.5	87.2
<i>Marginalized</i>	64.3	63.4	62.8	66.7	89.9
<i>Tanimoto</i>	66.4	64.2	65.7	66.9	87.8
<i>MiniMax</i>	64.0	64.5	64.5	66.4	86.2

Table 2. Results of classification on the datasets using subgraph based distances. The numbers are the classification accuracies in % using LOO cross validation.

much dependency of the performance on the radius on these datasets. This could be possibly due to the fact that atom label histograms match is a very weak sense throwing away all the information about the structure and as the radius grows the local histograms become more and more similar to each other, worsening the performance.

Our kernel performs considerably better (about 10% improvement) than the existing kernels for on all the 4 sets of the PTC Dataset and is comparable in the Mutag Dataset even for small radius. For $r = 1$ the subgraph based measures perform slightly better than the histogram based ones indicating that at a local level exact matching is useful. The subgraph based measures become intractable for higher radius so we have restricted our experiments for $r = 1$. A comparison of graph based and histogram based measures along with the others are shown in figure 5. The accuracies for the Mutag dataset do not show any improvement, suggesting that we need to look at higher order properties which might not necessarily be captured by labeled graphs.

kernel	MM	FM	MR	FR	Mutag
$r = 1$	74.3	72.3	72.5	78.2	87.8
$r = 2$	75.0	72.1	73.3	77.7	81.9
$r = 3$	75.0	72.0	73.3	78.2	84.6
$r = 4$	75.0	72.0	72.3	78.2	84.4

Table 3. Results of classification on the datasets using histogram measure for various context radius. The numbers are the classification accuracies in % using LOO cross validation.

6 Conclusion

In this work, we introduce a new distance between graphs with vertex and edge labels. We combine ideas from graph theory and Kernel Learning to use the metric to predict properties of chemicals molecules. The local nature of exact matching makes it computationally attractive while global properties are captured by matching scheme we propose. The derived kernels show improved performance over the existing kernel based techniques two well known datasets.

References

- [1] Helma, C., King, R., Kramer, S., Srinivasan, A.: Predictive toxicology challenge (2001)
- [2] Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)
- [3] Shapiro, L., Haralick, R.: Structural descriptions and inexact matching. (1981) 504–519
- [4] Bunke, H., Shearer, K.: A graph distance metric based on the maximal common subgraph (1998)
- [5] Le, S.Q., Ho, T.B., Phan, T.H.: A novel graph-based similarity measure for 2d chemical structures. *Genome Informatics* **15(2)** (2004) 82–91
- [6] Ralaivola, L., Swamidass, S.J., Saigo, H., Baldi, P.: Graph kernels for chemical informatics. *Neural Netw.* **18(8)** (2005) 1093–1110
- [7] Bennett, K.P., Cristianini, N., Shawe-Taylor, J., Wu, D.: Enlarging the margins in perceptron decision trees. *Mach. Learn.* **41(3)** (2000) 295–313
- [8] Gartner, T.: A survey of kernels for structured data. *SIGKDD Explor. Newsl.* **5(1)** (2003) 49–58
- [9] Gartner, T., Flach, P.A., Kowalczyk, A., Smola, A.J.: Multi-instance kernels. In Sammut, C., Hoffmann, A., eds.: *Proceedings of the 19th International Conference on Machine Learning*, Morgan Kaufmann (2002) 179–186
- [10] Kashima, H., Tsuda, K., Inokuchi, A.: Marginalized kernels between labeled graphs. In: *Proceedings of 20th International Conference of Machine Learning*. (2003)
- [11] Mahe, P., Ueda, N., Akutsu, T., Perret, J.L., Vert, J.P.: Extensions of marginalized graph kernels. *Machine learning* (2004)
- [12] Ramon, J., Bruynooghe, M.: A Polynomial Time Computable Metric between Point Sets, Technical Report CW301. Technical report (2000)
- [13] Scholkopf, B., Smola, A.J.: *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA (2001)
- [14] Debnath, de Compadre, A.L., R.L., Debnath, Shusterman, G., C, A.H.: Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds... correlation with molecular orbital energies and hydrophobicity. (1991)