Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(2):368-372

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Chemical compound classification based on improved Max-Min kernel

Qiangrong Jiang, Can Zhai and Zhikang Xiong

College of Computer Science and Technology, BJUT, Chaoyang District, Beijing No. 100 Pingleyuan, China

---

ABSTRACT

In this paper, a new method is defined that encapsulates the diameter information of the fingerprint. Considering fingerprints of the large diameter containing some large and quite precise information, such fingerprints should be given relative large weight. A review of the MIN–MAX kernel is provided followed by a thorough definition of the improved method by which the diameter information of the fingerprint was encoded. The paper concludes with comparative QSAR studies to test the efficacy of this improved MAX-MIN kernel in comparison with a number of published methods and results with two datasets: PTC and NCI for the convenience of comparison.

Keywords: ECFP, improved MAX-MIN kernel, PTC, NCI, classification accuracy.

INTRODUCTION

In the field of chemistry the term molecular fingerprint refers to the characterization of a molecular structure into a vector of descriptors or features. These descriptors can then be applied to a wide variety of problems in Chemo-informatics such as similarity searching [1], cluster analysis [2] and classification. Considering the wide used fingerprint, we give an introduction about the Hash-key fingerprints, which is the foundation of our implementation.

Hash-key fingerprints, although they also result in a vector-based (and typically binary) representation, have a distinctly alternative method of generation from structure key fingerprints. Each atom in a given molecule is iterated over, with all atom-bond paths being enumerated from that atom between a defined minimum and maximum bond path length. Each of these paths is then encoded using a Cyclic Redundancy Check (CRC) hashing algorithm into a single large integer, in the range $\{0...2^{32}–1\}$. This integer is then passed as the seed to a Random Number Generator (RNG), from which a defined number, N, of integers are taken. Each of these integers is then reduced into the length of the fingerprint, bits, by application of the modulus operator. This set of indices is then used to set or update the relevant positions in the fingerprint vector. The pseudo-code for the typical hash-key fingerprinting algorithm is provided here:

foreach *atom* **in** *molecule*
foreach *path* **from** *atom*
seed=**crc32**(*path*)
srand(*seed*)
for *i=1* **to** *N*
index=**rand**()%*bits*
setBit(*index*)

Hash-key fingerprints provide a rapid and efficient description of topological molecules. At the same time, it also has its own, somewhat complementary, limitations. Essentially, these limitations are characteristics of information-based methods when considering the hash-key fingerprints. Due to the method by which hash-key fingerprints are generated they are very difficult to interpret.

The rests of the paper are organized as follows. First of all, we introduced the representation of compounds, then, we give a detail interpretation of the ECFPs-based descriptor and its relationship with the Morgan Algorithm, we described the ECFP with a name-value method to adapt our kernel method, next, we described the improved Max-Min kernel method that we proposed and explained the advantage compared with the old method. Finally, we gave the experiment result and made contrast with the state-of-art method.

EXPERIMENTAL SECTION

Representation of compounds

In this paper we represent each compound by its corresponding molecular graph. The vertices of these graphs correspond to the various atoms (e.g., carbon, nitrogen, oxygen, etc.), and the edges correspond to the bonds between the atoms (e.g., single, double, etc.). Each of the vertices and edges has a label associated with it. The labels on the vertices correspond to the type of atoms and the labels on the edges correspond to the type of bonds. Specifically, we use a unique identifiers for each atomic number as the atom typing for vertices. For the edge labels, we use separate integers or identifiers for single, double and triple bonds. We also apply two commonly used structure normalization transformations. First, we label all bonds in aromatic rings as aromatic (i.e., a different edge-label), and second, we remove the hydrogen atoms that are connected to carbon atoms (i.e., hydrogen-suppressed chemical graphs). An atom numbering is shown in Fig. 1.

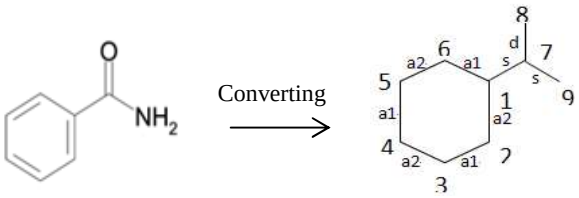

Fig.1 Benzoic acid amide atom numbering (of non-hydrogen atoms)

Extended-connectivity fingerprints (ECFPs)

ECFPs are a novel class of topological fingerprints for molecular characterization. Historically, topological fingerprints were developed for substructure and similarity searching but later used for analysis tasks, such as clustering, and classification. ECFPs are explicitly designed to capture molecular features relevant to molecular activity. While not designed for substructure searching, they are very suited to tasks related to predicting and gaining insight into drug activity.

ECFPs are circular fingerprints with a number of useful qualities: they can be very rapidly calculated; they are not predefined and can represent an essentially infinite number of different molecular features (including stereo-chemical information); their features represent the presence of particular substructures, allowing easier interpretation of analysis results; and the ECFP algorithm can be tailored to generate different types of circular fingerprints, optimized for different uses.

Relation to Morgan Algorithm

ECFPs are derived using a variant of the Morgan algorithm, In the Morgan algorithm, an iterative process assigns numeric identifiers to each atom, at first using a rule that encodes the numbering invariant atom information into an initial atom identifier, and later using the identifiers from the previous iteration. Thus, identifiers generated are independent of the original numbering of the atoms. The iteration process is continued until every atom identifier is unique (or as close to “unique” as symmetry allows); the intermediate results are discarded, and the final identifiers provide a canonical numbering scheme for the atoms.

The ECFP algorithm makes several changes to the standard Morgan algorithm. First, ECFP generation terminates after a predetermined number of iterations rather than after identifier uniqueness is achieved. The initial atom identifiers, and all identifiers produced each iteration are collected into a set; it is this set that defines the extended-connectivity fingerprint. Indeed, obtaining these partially disambiguated atom identifiers is the goal of the process. This means that the iteration process does not have to proceed to completion (that is, maximum disambiguation) but is performed for a predetermined number of iterations. Second, since perfectly accurate disambiguation is not required, algorithmic optimizations are possible. In the ECFP algorithm, this computationally expensive step is replaced by a fast hashing scheme. This results in a savings of computational effort when the ECFP algorithm is used for fingerprint generation, as compared to the rigorous Morgan algorithm used for canonicalization. Importantly, the ECFP-hashing scheme generates identifiers that are comparable across molecules. An example about ECFP is show as follows

The ECFP generation process has three sequential stages:
1. An initial assignment stage in which each atom has an integer identifier assigned to it.
2. An iterative updating stage in which each atom identifier is updated to reflect the identifiers of each atom's neighbors, including identification of whether it is a structural duplicate of other features.
3. A duplicate identifier removal stage in which multiple occurrences of the same feature are reduced to a single representative in the final feature list. (The occurrence count may be retained if one requires a set of counts rather than a standard binary fingerprint.)

Descriptor-based kernel functions.
Given the descriptor space, each chemical compound can be represented by a vector X whose i th dimension will have a non-zero value if the compound contains that descriptor and will have a value of zero otherwise. The value for each descriptor that is present can be either one, leading to a vector representation that captures presence or absence of the various descriptors (referred to as binary vectors) or the number of times (number of embedding) that each descriptor occurs in the compound, leading to a representation that also captures the frequency information (referred to as frequency vectors).

Given the above vector representation of the chemical compounds, the classification algorithms that we develop in this paper use support vector machines (SVM) as the underlying learning methodology, as they have been shown to be highly effective, especially in high dimensional spaces. One of the key parameters that affect the performance of SVM is the choice of the kernel function (K) that measures the similarity between pairs of compounds. Any function can be used as a kernel as long as, for any number n and any possible set of distinct compounds $\{X_1,...,X_n\}$,the n×n Gram matrix defined by $K_{i,j} = K(X_i, X_j)$ is symmetric positive semi-definite. These functions are said to satisfy Mercer's conditions and are called Mercer kernels, or simply valid kernels.

The Min–Max kernel [3] was selected because it has been shown to be an effective way to measure the similarity between chemical compound pairs and outperform Tanimoto coefficient [3] (which is the most widely used kernel function in chemo-informatics) in empirical evaluations. Given the vector representation of two compounds X and Y, the Min–Max kernel function is given by

$$\mathbf{K}_{MM}(X,Y) = \frac{\sum_{i=1}^{M} min(x_i,y_i)}{\sum_{i=1}^{M} max(x_i,y_i)} \quad (1)$$

where the terms x_i and y_i are the name-value mapping along the i th dimension of the X and Y vectors, respectively. Note that in the case of binary vectors, the value will be either zero or one, whereas in the case of frequency vectors the value will be equal to the number of times the i th descriptor exists in the two compounds. Moreover, note that the Min–Max kernel is a valid kernel as it has been shown to satisfy Mercer's conditions and reduces to Tanimoto kernel in the case of binary vectors.

One of the potential problems in using the above kernel with descriptor spaces that contain fingerprints of different diameter is that they contain no mechanism to ensure that descriptors of various diameters contribute in a non-trivial way to the computed kernel function values. This is especially true for the ECFP descriptor space in which the final set contains a mixture of fingerprints of differing diameters for each atom in the molecule, some large and quite precise should give relative large weight, while some small and relatively common should give relative small weight. To overcome this problem, we modified the above kernel function and give equal weight to the fingerprint of the same diameter. Particularly, for the Min–Max kernel function, this is obtained as follows. Let X^l and Y^l be the feature of X and Y with respect to only the fingerprints of diameter *l*, and let L be the diameter of the largest fingerprint. Then, the improved Max–Min kernel function $\mathbf{K}^{e}_{MM}(X,Y)$ is given by

$$\mathbf{K}^{e}_{MM}(X,Y) = \sum_{l=1}^{L} l * \mathbf{K}_{MM}(X^l,Y^l) \quad (2)$$

The construction of models is to give a set of molecules with some annotation of bioactivity. A wide variety of modeling methods are possible. The high dimensionality of ECFPs is a particular advantage for Bayesian analysis or Tanimoto (and related) similarty methods, as they make good use of the wide variety and large number of ECFP features.

ECFPs, being a topological method, do not directly represent 3D information. However, for many purposes, topological methods like ECFPs have advantages over 3D methods. In fact, 3D fingerprints are expensive to generate because of the need to generate 3D conformations, restricting their use to smaller data sets. The generation of

representative conformations is an area of ongoing research, and different conformational generation methods may result in vastly different 3D fingerprints. 3D fingerprints, such as affinity fingerprints, that reply on experimental data are also expensive to generate and are unavailable for virtual compounds.

Since the 3D conformation of molecules depends on the topological structure, topological information contains much of the same useful information as the 3D information. Indeed, in most published analyses of topological versus 3D descriptors, the authors came to the conclusion that topological descriptors are as good or superior to 3D descriptors for molecular tasks like similarity searching and activity prediction. There is, however, ongoing debate as to whether 3D fingerprints are better than topological fingerprints for "scaffold-hopping" between structural classes.

PTC dataset

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

Table 1 Distribution of positive and negative examples and molecular graph statistics in the PTC datasets

	MR	**FR**	**MM**	**FM**
No. of pos.	152 (44.2%)	121 (34.5%)	129 (38.4%)	143 (41.0%)
No. of neg.	192 (55.8%)	230 (65.5%)	207 (61.6%)	206 (59.0%)
Total ex.	344	351	336	349
Avg. no. of atoms/mol	25.56	26.08	25.05	25.25
Avg. no. of bonds/mol	25.96	26.53	25.39	25.62
Avg. degree	2.03	2.03	2.03	2.03

NCI dataset

In the NCI-HIV database, each compound is described by its chemical structure [4] and classified into three categories: confirmed inactive (CI), moderately active (CM), or active (CA). A compound is inactive if a test showed less than 50% protection of human CEM cells. All other compounds were retested. Compounds showing less than 50% protection (in the second test) are also classified inactive. The other compounds are classified active, if they provided 50% protection in both tests, and moderately active, otherwise. We formulated three problems out of this dataset. The first problem is designed to distinguish between CM+CA and CI; the second between CA and CI, and the third between CA and CM.

Comparison with published results

In recent years many new descriptors and graph kernels have been introduced in the data-mining literature and their classification performance has been successfully assessed. The performance assessment measure used in those studies is primarily area under the ROC curve. In Table 2 we compared the ROC results of ECFP with the results of Cycles and Trees (CT) [5], random-walk based graph kernels (RWK) [6], weighted decomposition kernels (WDK) [7] and Frequent sub-graph based descriptors [8]. We used the improved Max-Min kernel (K_{MM}) for ECFP-based descriptors. The results could only be compared for the common datasets with those used in these studies. We used the default misclassification cost factor (1.0) and did not optimize for regularization parameter in ECFP-based descriptors. We compared our result with the already reported result corresponding to the related descriptors, it can be observed from Table 2 that the ECFP-based descriptor outperforms CT, RWK, WDK and FSG for the majority of the datasets. Moreover, the best performing method consistently fell into ECFP descriptors (except CA vs. CM) despite the fact that no optimization performed on the SVM parameters. The average improvement of ECFP over the ROC values of WDK, CT, RWK, and FSG for the common datasets was 1.64, 1.72, 7 and 6.8%, respectively.

Table 2 ROC values for the five methods for chemical compound classification

Datasets	ECFP	CT	RWK	WDK	FSG
CA+CM vs CI	**0.832**	0.809		0.817	0.765
CA vs CI	**0.953**	0.925		0.94	0.839
CA vs CM	0.832	0.826		**0.842**	0.81
MR	**0.714**		0.632	0.697	0.626
FR	**0.685**		0.664	0.649	0.634
MM	**0.697**		0.656	0.705	0.655
FM	**0.758**		0.645	0.691	0.673

DISCUSSION AND CONCLUSION

The work in this paper was primarily motivated by our desire to understand which aspects of the molecular graph are

important in providing effective descriptor-based representations for the classification tasks given the four design choices (dataset specificity, fragment complexity, preciseness, and coverage) and the fact that no scheme leads to a descriptor space that is strictly superior (in terms of what it captures) to the rest of the schemes. Most of the descriptor spaces make some compromises along at least one of these dimensions. In the QSAR [9] model, we believe that the experimental results presented in Table 2 provide some answers on the relative importance and impact of these design choices.

Acknowledgements

This project is supported by Beijing Municipal Education Commission, the Grant number is 007000546313524. The authors are grateful to Beijing University of Technology for financial support.

REFERENCES

[1] P Willett; JM Barnard; GM Downs., *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 983 – 996.
[2] GM Downs; JM Barnard., *Rev. Comput. Chem.* **2002**, 18, 1 – 40.
[3] SJ Swamidass; J Chen; J Bruand; P Phung; L Ralaivola; P Baldi. *Bioinformatics*, **2005**, 21(1):359–368
[4] M Deshpande; M Kuramochi and G Karypis.. No. TR-02-027. Minnesota Univ Minneapolis Dept of Computer Science, **2002**.
[5] T Horváth; T Gärtner; S Wrobel. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, **2004**: 158-167.
[6] H Kashima; K Tsuda; A Inokuchi. *ICML*. **2003**, 3: 321-328.
[7] S Menchetti; F Costa; P Frasconi. Proceedings of the 22nd international conference on Machine learning. ACM, **2005**: 585-592.
[8] M Deshpande; M Kuramochi; N Wale et al. Knowledge and Data Engineering, IEEE Transactions on, **2005**, 17(8): 1036-1050.
[9] RK Prasad; T Narsinghani; R Sharma.,*Journal of Chemical and Pharmaceutical Research* v 1, n 1, p 199-206, **2009**