

## 1 实验研究

我们对我们的算法进行了广泛的实验研究来评估其有效性, 高效性及扩展性。我们在化学分子结构上测试我们的算法。对于化学结构, 节点特征包括数值特征和原子布尔特征。数值特征包括元素种类, 原子部分电荷, 原子电子亲和势, 原子自由电子数目和原子价态等等。布尔特征包括原子是否在供体中, 是否在末端碳中, 是否在环中, 是否为负, 是否是轴向的等等。在实验中, 我们仅用一个原子特征: 元素种类。

我们将我们的方法和小波分配核, C-tree, GraphGrep 还有 gIndex 进行对比。我们的算法, WA 算法, GraphGrep 和 gIndex 是基于 C++ 实现的, 用 g++ 进行编译。C-tree 是用 Java 实现的, 用 Sun JDK 1.5.0 编译。所有的实验都是在 Intel Xeon EM64T 3.2GHz, 4G 内存, Linux 系统这一平台上测试的。

WA, G-Hash, C-tree, GraphGrep 和 gIndex 的参数是这样设置的。对于 WA 和 G-hash,  $h$  取 2, 用 *haar* 小波函数, 对于 C-tree, 用默认值即将最小子节点数  $m$  设为 20, 最大  $M$  设为  $2m - 1$ , 用 NBM 方法进行图映射。对于 GraphGrep 和 gIndex, 全部采用默认参数。

### 1.1 数据集

我们选用许多数据集来进行试验。前五个数据集是从 Jorissen/Gilson 数据集获得的已有数据。接下来六个是从 BindingDB 数据集中抽取的, 最后一个 NCI/NIH 艾滋病筛选集里的, 表 1 显示了这些数据集和其基本情况。

#### 1.1.1 Jorissen 数据集

Jorissen 数据集主要为一些包含活动蛋白质的化学结构信息。我们以药物对特定蛋白质的吸引力作为目标值。我们选取了 5 个包含 100 个分子结构的蛋白质作为测试目标, 其中 50 个化学结构连接着蛋白质, 另外 50 个没连接到蛋白质上。参考文献 14 来查看详细信息。

#### 1.1.2 NCI/NIH AIDS 抗体数据库

NCI/NIH 艾滋病抗体数据库包含 42390 个化学结构, 总共有 63 种分子, 最常见的是, C, O, N, S。数据集包括三种边, 单边, 双边, 芳香边。我们选取了所有结构作为数据库, 并随机选取 1000 个化学结构作为查询集合。

### 1.2 利用分类做节点相似性度量