



پردیس بین‌المللی ارس
گروه مهندسی کامپیوتر

پایان‌نامه

برای دریافت درجه کارشناسی‌ارشد در رشته مهندسی کامپیوتر گرایش نرم‌افزار

عنوان

بررسی مقایسه‌ای کاربرد روش‌های انتخاب پارامتر در مطالعات QSAR

استاد راهنما

دکتر محمدعلی بالافر

استاد مشاور

دکتر سمیه سلطانی

پژوهشگر

سیامک زنجانی

شهریور ۱۳۹۳

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

سپاسگزاری

در ابتدا از زحمات استاد راهنما آقای دکتر محمد علی بالافر و استاد مشاور خانم دکتر سمیه سلطانی تشکر می‌نمایم، همچنین از دوستان گرامی خانم‌ها محمدی‌وند، نورزاده و آقایان ذکریازاده و سعیدوند که مرا با راهنمایی‌های خودشان یاری کردند تشکر می‌کنم. در پایان از زحمات پدر و مادر و خانواده خودم تشکر می‌کنم.

سیامک زنجانی

۱۳۹۳

نام خانوادگی دانشجو : زنجانی نام : سیامک

عنوان پایان نامه: بررسی مقایسه‌ای کاربرد روش‌های انتخاب پارامتر در مطالعات QSAR

استاد راهنما: آقای دکتر محمد علی بالافر
استاد مشاور: خانم دکتر سمیه سلطانی

مقطع تحصیلی: کارشناسی ارشد رشته: مهندسی کامپیوتر گرایش: نرم افزار دانشگاه: تبریز
دانشکده: پردیس بین‌المللی ارس تاریخ فارغ‌التحصیلی: ۱۳۹۳/۶/۲۰ تعداد صفحه: ۱۳۷

کلید واژه: انتخاب ویژگی، متلب، جعبه ابزار، استخراج داده، بررسی کمی رابطه‌ها ساختمان-فعالیت ترکیب‌ها

چکیده:

QSAR یا بررسی رابطه‌های کمی ساختمان و فعالیت ترکیب‌ها روشی است که مدل‌های ریاضی یا کامپیوتری را به منظور یافتن رابطه معنی‌دار آماری بین فعالیت و ساختار بکار می‌گیرد. یکی از مهمترین مرحله‌های انجام QSAR انتخاب ویژگی است. انتخاب ویژگی عبارت است از انتخاب یک زیر مجموعه از ویژگی‌های مجموعه داده بطوری که این ویژگی‌ها اطلاعات مفیدی در اختیار ما بگذارند. فایده‌های انتخاب ویژگی عبارت است از کاهش زمان محاسبه، حافظه مورد نیاز، کاهش زمان یادگیری، بهبود کارایی پیشگویی و حذف داده‌های مخدوش است. ما انواع روش‌های انتخاب ویژگی که شامل روش‌های خطی و غیرخطی است را بکار بردیم. برای انجام انتخاب ویژگی از جعبه‌ابزارهای مختلفی استفاده کرده‌ایم. برای مدل‌سازی از دو روش استفاده شده است: رگرسیون مرحله‌ای که یک روش خطی است و شبکه عصبی که روش غیرخطی محسوب می‌گردد. به منظور انجام سریع و یکپارچه محاسبات 2D-QSAR، جعبه‌ابزاری را در متلب ایجاد کرده‌ایم. نام این جعبه ابزار FQSAR است. از مزیت‌های این جعبه‌ابزار افزایش سرعت و دقت در انجام محاسبه‌ها است. به راحتی می‌توان انواع روش‌های انتخاب ویژگی و مدل‌ها را به آن اضافه نمود. برای ارزیابی جعبه‌ابزار و روش‌های انتخاب ویژگی، نتیجه‌ی محاسبه‌ها بر روی سه تا مجموعه داده با استفاده از این جعبه ابزار مورد بررسی قرار گرفته است.

فهرست مطالب

۱	فصل اول - معرفی
۲	۱-۱ انگیزه
۲	۲-۱ بیان مساله
۴	۳-۱ اهداف
۴	۴-۱ چهارچوب پایان نامه
۵	فصل دوم - بررسی منابع
۶	۱-۲ QSAR چیست؟
۶	۱-۱-۲ ۱-۱-۲ توصیفگرها
۷	۲-۲-۱ انواع روش‌های QSAR
۷	۱-۱-۲-۱-۲ 1D-QSAR
۷	۲-۲-۱-۱ 2D-QSAR
۸	۳-۲-۱-۲ 3D-QSAR
۸	۴-۲-۱-۲ 4D-QSAR روش
۹	۵-۲-۱-۲ 5D-QSAR روش
۹	۶-۲-۱-۲ 6D-QSAR روش
۹	۳-۱-۲ مرحله‌های مختلف ارزیابی مدل QSAR
۱۱	۴-۱-۲ پایگاه داده‌های مورد استفاده در مطالعه‌ها QSAR
۱۲	۵-۱-۲ نرم افزارهای مورد استفاده در مطالعه‌ها QSAR
۱۳	۲-۲ انتخاب توصیف‌گر (متغیر مستقل)
۱۴	۱-۲-۲ انواع FS
۱۵	۲-۲-۲ روش‌های رپر
۱۵	۳-۱-۲-۲ روش‌های تعبیه شده
۱۶	۴-۱-۲-۲ روش‌های رهبری شده و نشده FS
۱۶	۲-۲-۲ الگوریتم‌های انتخاب ویژگی
۱۷	۱-۲-۲-۲ تعاریف

۱۷.....	CFS الگوریتم ۲-۲-۲-۲
۱۸.....	Chi-Square [۱۵] الگوریتم ۳-۲-۲-۲
۱۹.....	CIFE الگوریتم ۴-۲-۲-۲
۱۹.....	DISR الگوریتم ۵-۲-۲-۲
۲۰.....	FAS الگوریتم ۶-۲-۲-۲
۲۱.....	FCBF الگوریتم ۷-۲-۲-۲
۲۲.....	Information Gain الگوریتم ۸-۲-۲-۲
۲۲.....	Gini Index الگوریتم ۹-۲-۲-۲
۲۲.....	HSIC الگوریتم ۱۰-۲-۲-۲
۲۳.....	ICAP الگوریتم ۱۱-۲-۲-۲
۲۴.....	JMI الگوریتم ۱۲-۲-۲-۲
۲۴.....	KFDS الگوریتم ۱۳-۲-۲-۲
۲۵.....	MIFS الگوریتم ۱۴-۲-۲-۲
۲۵.....	MRMR الگوریتم ۱۵-۲-۲-۲
۲۶.....	Relief الگوریتم ۱۶-۲-۲-۲
۲۶.....	SAS الگوریتم ۱۷-۲-۲-۲
۲۷.....	SMBLR الگوریتم ۱۸-۲-۲-۲
۲۸.....	SFS الگوریتم ۱۹-۲-۲-۲
۲۸.....	F-score, t-score الگوریتم‌های ۲۰-۲-۲-۲
۲۸.....	کاهش بعد ۳-۲
۲۸.....	LPP الگوریتم ۱-۳-۲
۲۹.....	NPE ^۴ الگوریتم ۲-۳-۲
۳۰.....	NCA الگوریتم ۳-۳-۲
۳۲.....	فصل سوم - روش‌ها
۳۳.....	۱-۳ ابداع مدل QSAR

۳۵	۲-۳ معتبرسازی ^۱ مدل
۳۵	۱-۲-۳ محاسبه ضریب رگرسیون (R^2)
۳۵	۲-۲-۳ leave one out cross validation (LOOCV)
۳۶	۳-۳ محاسبه خطای آموزش و یادگیری
۳۶	۴-۳ آزمون تصادفی بودن مدل
۳۶	۵-۳ توسعه جعبه ابزار
۳۹	۱-۵-۳ الگوریتم‌های انتخاب متغیر وارد شده در FQSAR
۴۰	۲-۵-۳ تابع‌های مورد استفاده برای توسعه FQSAR
۴۰	۱-۲-۵-۳ تابع Zeron
۴۰	۲-۲-۵-۳ تابع outlayer
۴۱	۳-۲-۵-۳ تابع dataprepar
۴۱	۶-۵-۲-۳ تابع IntercorrelationR2
۴۲	۷-۵-۲-۳ تابع findfeature
۴۲	۶-۲-۵-۳ تابع FS
۴۲	۷-۲-۵-۳ تابع RunFS
۴۳	۸-۲-۵-۳ تابع ANN1
۴۳	۹-۲-۵-۳ تابع Modeling
۴۴	۱۰-۲-۵-۳ تابع RandomchangeExp1
۴۴	۱۱-۲-۵-۳ تابع Randomcheckmodel
۴۵	۱۲-۲-۵-۳ تابع q2
۴۵	۱۳-۲-۵-۳ تابع q2ANN1
۴۶	۱۴-۲-۵-۳ تابع initialdatatotal...
۴۶	۱۵-۲-۵-۳ تابع SaveSignifitiont
۴۶	۱۶-۲-۵-۳ تابع creat1
۴۶	۱۷-۲-۵-۳ تابع Outputprepar
۴۶	۱۸-۲-۵-۳ تابع saveexceloutput
۴۷	۱۹-۲-۵-۳ فایل QSAR
۴۷	۲۰-۲-۵-۳ فایل prepar

۴۷.....	Export فایل ۲۱-۲-۵-۳
۴۷.....	graph1 فایل ۲۲-۲-۵-۳
۴۷.....	SelectedFS فایل ۲۳-۲-۵-۳
۴۸.....	فصل چهارم - نتایج.....
۴۹.....	۱-۴ یکپارچه‌سازی انجام شده توسط FQSAR.....
۴۹.....	۲-۴ نتایج‌های بدست آمده
۵۰.....	۱-۲-۴ نتایج‌های بدست آمده بر روی ترکیب‌های آنتاگونیست گیرنده آنژیوتانسین (ANG).....
۵۰.....	۱-۲-۴-۱ نتایج‌ها براساس خطا.....
۵۰.....	۱-۲-۴-۱-۱ آزمون تصادفی بودن مدل‌ها.....
۵۱.....	۲-۴-۱-۱ تنظیم‌های اولیه جعبه ابزار برای انجام محاسبه‌های
۵۲.....	۲-۴-۱-۱ پارامترها و ضریب‌های انتخاب شده.....
۵۴.....	۲-۴-۱-۱ نتایج‌ها محاسبه‌های براساس خطا
۵۶.....	۲-۴-۱-۱ مقایسه نتایج‌ها بدست آمده بر اساس تعداد ویژگی‌های انتخاب شده
۵۷.....	۲-۴-۱-۲ نتایج‌ها براساس R2.....
۵۷.....	۲-۴-۱-۲ آزمون تصادفی بودن مدل‌ها.....
۵۹.....	۲-۴-۱-۲ تنظیم‌های اولیه جعبه ابزار برای انجام محاسبه‌ها
۵۹.....	۲-۴-۱-۲ پارامترها و ضریب‌های انتخاب شده.....
۶۱.....	۲-۴-۱-۲ نتایج‌ها براساس R2.....
۶۳.....	۲-۴-۱-۲ مقایسه نتایج بدست آمده بر اساس تعداد ویژگی‌های انتخاب شده
۶۴.....	۲-۴-۱-۲ نتایج‌های بدست آمده بر روی ترکیب‌های MPr.....
۶۴.....	۲-۴-۱-۲ نتایج‌ها براساس خطا.....
۶۴.....	۲-۴-۱-۲ آزمون تصادفی بودن مدل‌ها.....
۶۵.....	۲-۴-۱-۲ تنظیم‌های اولیه جعبه ابزار برای انجام محاسبه‌ها
۶۶.....	۲-۴-۱-۲ پارامترها و ضریب‌های انتخاب شده.....
۶۹.....	۲-۴-۱-۲ نتایج‌ها براساس خطا.....
۷۰.....	۲-۴-۱-۲ مقایسه نتایج‌های بدست آمده بر اساس تعداد ویژگی‌های انتخاب شده
۷۱.....	۲-۴-۱-۲ نتایج‌ها براساس R2.....
۷۲.....	۲-۴-۱-۲ آزمون تصادفی بودن مدل‌ها.....

۷۳	تنظیم‌های اولیه جعبه ابزار برای انجام محاسبه‌ها.....
۷۳	پارامترها و ضریب‌های انتخاب شده.....
۷۷	نتیجه‌ها براساس R2.....
۷۸	مقایسه نتیجه‌های بدست آمده بر اساس تعداد ویژگی‌های انتخاب شده
۸۱	نتیجه‌های بدست آمده بر روی ترکیب‌های MPT.....
۸۱	نتیجه‌ها براساس خطا.....
۸۱	آزمون تصادفی بودن مدل‌ها.....
۸۲	تنظیم‌های اولیه جعبه ابزار برای انجام محاسبه‌ها
۸۲	نتیجه‌های بدست آمده براساس خطا
۸۴	مقایسه نتیجه‌های بدست آمده بر اساس تعداد ویژگی‌های انتخاب شده
۸۴	نتیجه‌ها براساس R2
۸۴	آزمون تصادفی بودن مدل‌ها.....
۸۶	تنظیم‌های اولیه جعبه ابزار برای انجام محاسبه‌ها
۸۶	نتیجه‌های بدست آمده براساس R2
۸۷	مقایسه نتیجه‌های بدست آمده بر اساس تعداد ویژگی‌های انتخاب شده و R2
۸۸	پارامترها و ضریب‌های انتخاب شده.....
۱۰۰	فصل پنجم - بحث، نتیجه گیری و پیشنهادات
۱۰۱	۱-۵ بحث و نتیجه‌گیری
۱۰۱	۲-۵ پیشنهادات
۱۰۳	پیوست الف -
۱۰۳	جدول‌های نتیجه‌ها
۱۰۴	پیوست الف - ۱ ماتریس ارتباط‌های الگوریتم FASrbf مجموعه داده ANG
۱۰۴	پیوست الف - ۲ ماتریس ارتباط‌های الگوریتم KDFS مجموعه داده ANG
۱۰۴	پیوست الف - ۳ ماتریس ارتباط‌های الگوریتم SCSS مجموعه داده MPr
۱۰۵	پیوست الف - ۴ ماتریس ارتباط‌های الگوریتم CMIM مجموعه داده MPr
۱۰۶	پیوست ب -
۱۰۶	راهنمای استفاده از برنامه
۱۰۷	پیوست ب-۱ روش اجرای برنامه

۱۰۷.....	پیوست ب-۲ راهنمای اجرای برنامه
۱۰۹.....	پیوست ب-۲-۱ بخش مربوط به FEAST
۱۱۰.....	پیوست ب-۲-۲ بخش مربوط به FEAT
۱۱۰.....	پیوست ب-۲-۳ بخش مربوط به ADR
۱۱۱.....	پیوست ب-۲-۴ بخش مربوط به HSICLASSO
۱۱۲.....	پیوست ب-۲-۵ بخش مربوط به FS
۱۱۳.....	پیوست ب-۲-۶ بخش مربوط به Stepwise
۱۱۳.....	پیوست ب-۲-۷ بخش مربوط به GA
۱۱۳.....	پیوست ب-۳ خروجی برنامه
۱۱۳.....	پیوست ب-۳-۱ شیت Value
۱۱۵.....	پیوست ب-۳-۲ شیت Coef
۱۱۵.....	پیوست ب-۳-۳ شیت Y randomization
۱۱۵.....	پیوست ب-۳-۵ شیت Bivariate Correlation
۱۱۵.....	پیوست ب-۳-۶ شیت Significant Value
۱۱۶.....	پیوست ب-۳-۷ شیت Significant Coef
۱۱۶.....	پیوست ب-۴ نمودارها
۱۱۶.....	پیوست ب-۵ Export
۱۱۸.....	منابع و مراجع

فهرست شکل‌ها

- شکل ۱-۱: مرحله‌های کلی انجام QSAR ۳
- شکل ۱-۲: رابطه بین فعالیت و ساختار ۶
- شکل ۲-۲: مرحله‌های انجام RD-4D-QSAR ۹
- شکل ۴-۲: الگوریتم SAS ۲۷
- شکل ۱-۳: مرحله‌های کلی اجرای جعبه ابزار ۳۷
- شکل ۱-۴: یکپارچه‌سازی عملیات QSAR ۴۹
- شکل ۲-۴: نمودارهای رابطه بین خطا و تعداد ویژگی‌های انتخاب شده ۵۷
- شکل ۳-۴: نمودارهای مقایسه R2 و تعداد ویژگی‌های انتخاب شده ۶۴
- شکل ۴-۴: نمودارهای رابطه بین خطا و تعداد ویژگی‌های انتخاب شده ۷۱
- شکل ۵-۴: نمودارهای مقایسه ای تعداد ویژگی انتخاب شده با R2 ۸۰
- شکل ۶-۴: نمودارهای مقایسه‌ای تعداد ویژگی انتخاب شده با R2 ۸۰
- شکل ۷-۴: نمودارهای مقایسه‌ای تعداد ویژگی انتخاب شده با خطا ۸۴
- شکل ۸-۴: نمودارهای مقایسه تعداد ویژگی‌های انتخاب شده با R2 ۸۸
- شکل پیوست ب-۱: شکل فرم ورد اطلاعات مربوط به آماده‌سازی داده‌ها و داده‌های اولیه ۱۰۸
- شکل پیوست ب-۲: خروجی فرم آماده‌سازی داده‌ها ۱۰۹
- شکل پیوست ب-۳: اجرای برنامه در FEAST ۱۱۰
- شکل پیوست ب-۴: اجرای ADR ۱۱۱
- شکل پیوست ب-۶: اجرای بخش FS ۱۱۳
- شکل پیوست ب-۷: لیست داده‌های که در صفحه Value آورده شده است ۱۱۴
- شکل پیوست ب-۸: لیست مربوط به شیت Coef ۱۱۵
- شکل پیوست ب-۱۰: شیت Bivariate Correlation ۱۱۵
- شکل پیوست ب-۱۱: شکل مربوط به نمودارهای مقایسه‌ای ۱۱۶

شکل پیوست ب-۱۲: فرم مربوط به استخراج ویژگی‌ها ۱۱۷

فهرست جدول‌ها

- جدول ۱-۲: لیست پایگاه داده‌های مربوط به QSAR ۱۱
- جدول ۲-۲: لیست نرم‌افزار بکار رفته در QSAR ۱۲
- جدول ۳-۲: لیست الگوریتم‌های انتخاب ویژگی ۱۵
- جدول ۱-۳: لیست الگوریتم‌های انتخاب ویژگی ۳۹
- جدول ۱-۴: توضیح‌های ستون‌ها ۵۰
- جدول ۲-۴: لیست مقدارهای تست تصادفی بودن مدل برای مجموعه داده ANG ۵۱
- جدول ۳-۴: تنظیم‌های اولیه برنامه در مجموعه داده ANG ۵۲
- جدول ۴-۴: لیست پارامترهای انتخاب شده ۵۲
- جدول ۵-۴: لیست نتیجه‌های محاسبه‌ها ۵۵
- جدول ۶-۴: لیست مقدارهای تست تصادفی بودن مدل برای مجموعه داده ANG ۵۸
- جدول ۷-۴: لیست پارامترهای انتخاب شده ۵۹
- جدول ۸-۴: لیست نتیجه‌های محاسبه‌ها ۶۲
- جدول ۹-۴: لیست مقدارهای تست تصادفی بودن مدل برای مجموعه داده MPr ۶۴
- جدول ۱۰-۴: تنظیم‌های اولیه برنامه ۶۵
- جدول ۱۱-۴: لیست پارامترهای انتخاب شده ۶۶
- جدول ۱۲-۴: لیست نتیجه‌های محاسبه‌ها ۶۹
- جدول ۱۳-۴: لیست مقدارهای تست تصادفی بودن مدل برای مجموعه داده MPr ۷۲
- جدول ۱۴-۴: لیست پارامترهای انتخاب شده ۷۳
- جدول ۱۵-۴: لیست نتیجه‌های محاسبه‌ها ۷۸
- جدول ۱۶-۴: لیست مقدارهای تست تصادفی بودن مدل برای مجموعه داده MPt ۸۱
- جدول ۱۷-۴: تنظیم‌های اولیه برنامه ۸۲
- جدول ۱۸-۴: نتیجه‌های محاسبه‌ها براساس خطا ۸۳

- جدول ۴-۱۹: لیست مقدرهای تست تصادفی بودن مدل برای مجموعه داده MPT ۸۵
- جدول ۴-۲۰: نتیجه‌ها محاسبه‌ها براساس R2 ۸۶
- جدول ۴-۲۱: لیست پارامترهای انتخاب شده ۸۸
- جدول ۴-۲۲: لیست پارامترهای انتخاب شده ۹۱
- جدول ۴-۲۳: لیست پارامترهای انتخاب شده ۹۴
- جدول ۴-۲۴: لیست پارامترهای انتخاب شده ۹۶
- جدول ۴-۲۵: لیست پارامترهای انتخاب شده ۹۷
- جدول ۴-۲۶: لیست پارامترهای انتخاب شده ۹۸
- جدول پیوست الف-۱: نتیجه ماتریس ارتباطی FASrbf ۱۰۴
- جدول پیوست الف-۲: نتیجه ماتریس ارتباطی KDFS ۱۰۴
- جدول پیوست الف-۳: نتیجه ماتریس ارتباطی SCSS ۱۰۴
- جدول پیوست الف-۴: نتیجه ماتریس ارتباطی CMIM ۱۰۵
- جدول پیوست ب-۱: فیلدهای مربوط به Stepwise ۱۱۴
- جدول پیوست ب-۲: لیست مربوط به شبکه عصبی ۱۱۴

فصل اول - معرفی

QSAR یا بررسی رابطه‌های کمی ساختمان و فعالیت ترکیب‌ها روشی است که مدل‌های ریاضی یا کامپیوتری را به منظور یافتن رابطه معنی‌دار آماری بین فعالیت^۱ و ساختار^۲ بکار می‌گیرد. این روش برای پیش‌بینی فعالیت شیمیایی ترکیب‌های جدید بکار می‌رود.

۱-۱ انگیزه

امروزه بیش از ده‌ها میلیون ترکیب^۳ در دنیا وجود دارد که هر کدام از آنها دارای خاصیت‌های مخصوص به خود است و این تعداد دائما در حال افزایش است. سه روش برای مطالعه فعالیت زیست‌شناسی ترکیب‌های شیمیایی وجود دارد: "in vivo" بکارگیری موجودهای زنده (مثل حیوان‌های آزمایشگاهی)، "in vitro" به کارگیری سلول‌ها، بیومولکول‌ها، ... و "in silico" که شامل به کارگیری روش‌های محاسبه‌ای برای پیش‌بینی فعالیت زیستی و یا خصوصیت‌های فیزیکی-شیمیایی ترکیب‌ها است.

روش‌های تجربی دارای عیب‌هایی هستند، از جمله اینکه هزینه و زمان زیادی برای اجرای آن‌ها نیاز است و همچنین غیرممکن بودن انجام برخی آزمایش‌ها بر روی انسان‌ها و یا حیوان‌ها محدودیت‌هایی را برای پژوهشگران ایجاد می‌کند. با توجه به مطلب‌های ذکر شده و نیز افزایش داده‌ها و اطلاعات در دسترس پژوهشگران، و در کنار همه این‌ها با افزایش قدرت پردازش کامپیوترها برای انجام محاسبه‌های پیچیده و حجیم، امروزه روش‌های "in silico" به عنوان ابزار کمکی و یا حتی جایگزین روش‌های تجربی برای پیش‌بینی فعالیت زیستی ترکیب‌ها مورد استقبال پژوهشگران قرار گرفته است. یکی از معمول‌ترین روش‌های مورد استفاده برای پیش‌بینی فعالیت ترکیب‌ها استفاده از رابطه‌های کمی ساختمان فعالیت (QSAR) است. QSAR به روش‌هایی اطلاق می‌شود که از یک رابطه کمی مبتنی بر مشخصه‌های ساختمانی ترکیب‌ها برای پیش‌بینی فعالیت آن‌ها استفاده می‌شود.

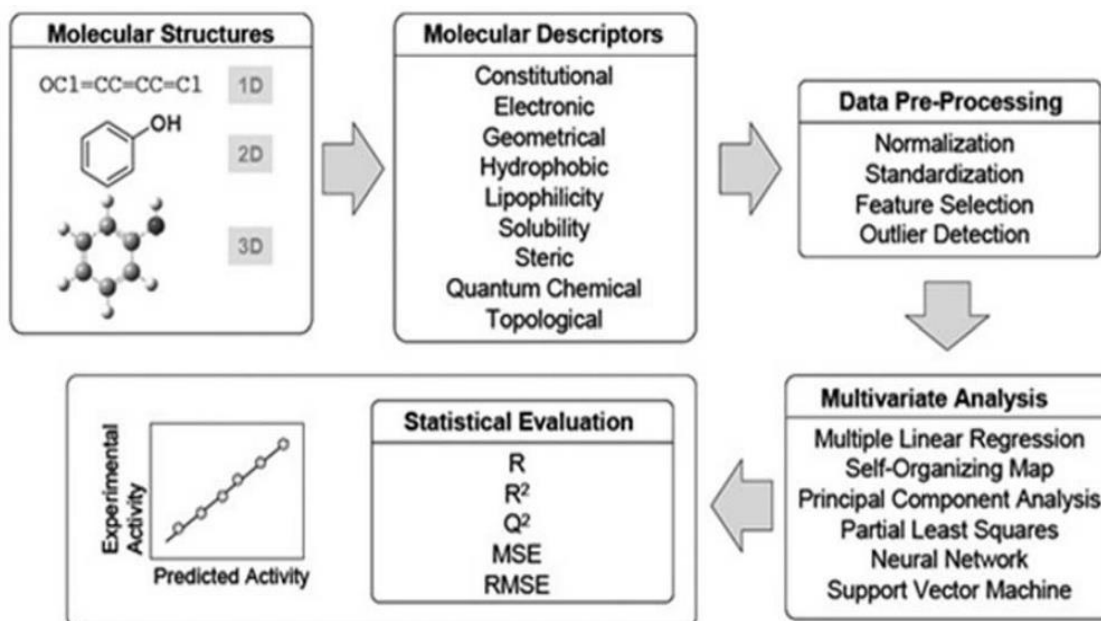
مزیت‌های روش QSAR به شرح زیر است:

- امکان پیش‌بینی ویژگی‌های فیزیکی - شیمیایی و فعالیت زیستی مولکول‌ها.
- کاهش هزینه مراحل کشف و توسعه داروها.
- امکان پیش‌بینی اثرهای جانبی داروها قبل از تولید انبوه.

۲-۱ بیان مساله

1 Activity
2 Structure
3 Compounds

- به طور کلی انجام QSAR را در شش مرحله کلی به شرح زیر می‌توان خلاصه کرد:
- تهیه پایگاه داده‌ها (گردآوری داده‌های مربوط به فعالیت زیستی یا خصوصیت‌های فیزیکی- شیمیایی)
 - محاسبه و یا تعیین توصیفگرهای ساختمانی.
 - آماده‌سازی داده‌ها^۱ شامل: نرمال‌سازی^۲ داده‌ها، استانداردسازی، انتخاب توصیفگرهای^۳ مناسب و حذف داده‌های پرت
 - ابداع رابطه کمی (مدل) مناسب بین توصیفگرها و فعالیت
 - معتبرسازی مدل ارایه شده
 - بازگشت به مرحله‌های قبلی و اعمال تغییرات مورد نیاز در صورت عدم موفقیت مدل



شکل ۱-۱: مرحله‌های کلی انجام QSAR

یکی از مهمترین مرحله‌های انجام QSAR مرحله انتخاب ویژگی است. این موفقیت منوط به انتخاب توصیفگرهای مناسب است که شامل حذف بعضی از توصیفگرهای نامرتب (مخدوش و افزونگی) و یا انتخاب پارامترهای با قدرت تبیین بالاتر فعالیت مورد مطالعه است.

1 Data Pre-Processing
2 Normalization
3 Descriptors

در این مطالعه روش‌های مختلف انتخاب توصیفگر از منابع استخراج شده و کاربرد آن‌ها در مطالعات QSAR مورد مطالعه قرار خواهد گرفت. بدین منظور جعبه‌ابزار مناسبی توسط نرم‌افزار متلب طراحی شده و مورد استفاده قرار خواهد گرفت. کارآیی روش‌های مختلف با استفاده از شاخص‌های معتبرسازی استخراج شده از دستورالعمل‌های بین‌المللی مورد بررسی قرار گرفته و مدل‌های مناسب ابداع خواهد شد.

۳-۱ اهداف

هدف اصلی این پایان‌نامه، مطالعه و مقایسه کاربرد انواع روش‌های انتخاب ویژگی در QSAR است. برای رسیدن به این هدف جعبه‌ابزاری^۱ با ویژگی‌های ذیل طراحی خواهد شد.

- کلیه داده‌های خروجی را به صورت یک فایل اکسل در دسترس کاربر قرار دهد.
- یکپارچگی: نرم‌افزاری که با یک عملیات ساده بتوان کلیه مرحله‌های ابداع و معتبرسازی یک مدل QSAR را انجام داد.
- امکان افزودن الگوریتم‌های جدید به سادگی میسر است.

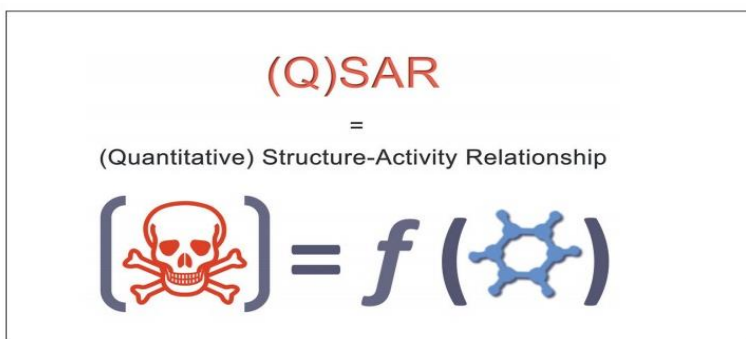
۴-۱ چهارچوب پایان‌نامه

در فصل دوم ابتدا به معرفی انواع روش‌های QSAR پرداخته و سپس نرم‌افزارهای بکار رفته به طور مختصر توضیح داده می‌شود، در ادامه فصل دوم در مورد انتخاب ویژگی بحث و سپس روش‌های مختلف آن مورد بررسی قرار می‌گیرد، در فصل سوم روش بکار رفته در محاسبه‌ها توضیح داده خواهد شد، و سپس کدهای جعبه‌ابزار مورد تجزیه و تحلیل قرار می‌گیرد. در فصل چهارم نتیجه‌های بدست آمده بر روی مجموعه داده‌ها بیان خواهد شد و در فصل پنجم پس از جمع‌بندی نتیجه‌ها، پیشنهادهایی برای مطالعه‌های آتی ارائه خواهد شد. در پیوست ب راهنمای اجرای برنامه آورده شده است.

فصل دوم - بررسی منابع

۱-۲ QSAR چیست؟

QSAR به رابطه‌های کمی بین فعالیت و یا خصوصیت‌های فیزیکی- شیمیایی یک ترکیب با خصوصیت‌های ساختمانی آن اطلاق می‌شود (شکل ۱-۲). روش‌های مدرن QSAR برای اولین بار در سال ۱۹۶۴ توسط فوجیتا^۱، هانس^۲ [۱] ویلسون^۳ ابداع گردید و پس از آن به سرعت گسترش یافت.



شکل ۱-۲: رابطه بین فعالیت و ساختار

۱-۱-۲ توصیفگرها

توصیفگرها در واقع مشخص کننده ویژگی‌های مولکول هستند، به دو دسته زیر تقسیم می‌گردند:

• توصیفگرهای دوبعدی (2D)

این نوع توصیفگرها در واقع توصیفگرهای توپولوژیکی گفته می‌شود، که نیاز به نمایش مسطح^۴ مولکول دارند مانند تعداد اتم‌های مولکول.

• توصیفگرهای سه بعدی (3D)

گونه دیگری از توصیفگرها نیاز به نمایش سه بعدی^۵ دارند که به آن‌ها توصیفگرهای سه بعدی گفته می‌شود، از قبیل توصیفگرهایی که مربوط به زاویه بین اتم‌های مولکول و غیره است. مقدار توصیفگرهای سه بعدی بسته به نوع نرم‌افزار، کاربر و تقریب‌هایی که بکار

1 Fujita
2 Hansch
3 Wilson
4 3D descriptors
5 Flat