

«طرح پیشنهادی پایان نامه کارشناسی ارشد»

۱- مشخصات دانشجو:

نام و نام خانوادگی : محمدمین خدامرادی	شماره دانشجویی : ۹۵۴۲۲۰۶۴
سال ورود : 95-96	رشته تحصیلی : علوم کامپیوتر
گرایش : بدون گرایش	تعداد واحدهای گذرانده شده : ۱۶
شماره تلفن برای تماس ضروری: ۰۹۲۲۷۱۳۶۰۵۱	
آدرس پست الکترونیکی: khodamoradi1992@gmail.com	

۲- نام و نام خانوادگی استاد (استادان) راهنما:

دکتر چنگیز اصلاحچی

۳- اطلاعات مربوط به پایان نامه:

۳-۱- تعداد واحد: ۶

۳-۲- عنوان:

فارسی : مقایسه‌ی روش‌های پیش‌بینی تعامل دارو و هدف

انگلیسی : Comparison of drug-target interaction prediction methods

۳-۳- واژگان کلیدی:

فارسی: تعامل دارو و هدف، استنتاج مبتنی بر شبکه‌ی گراف، آزمون AUPR، آزمون AUC، اعتبارسنجی متقابل

انگلیسی: Drug Target Interaction, network-based inference, AUPR test, AUC test, Cross validation

☐ توسعه ای☒ کاربردی۳-۴- نوع پایان نامه: بنیادی ☐

۳-۵- تعریف مسأله یا بیان موضوع مطالعه (سابقه، ضرورت و اهداف پایان نامه):

زمان و هزینه دو عامل مهم در صنعت داروسازی است. به طور میانگین، حدود پانزده سال زمان و هشتصد میلیون دلار هزینه نیاز است تا یک دارو بتواند در این سال‌ها جایگاه خود را در داروخانه‌ها پیدا کند. به علاوه حدود ۹۰ درصد داروها، در همان مراحل نخست کارآزمایی بالینی با شکست مواجه می‌شوند. مهم‌ترین دلایل شکست خوردن دارو ناکارآمدی و عوارض جانبی آن است. اخیراً دانشمندان در پی آن هستند سه مسئله‌ی مهم در این زمینه را تاحدی حل کنند:

۱. برای داروهای موجود اهداف دیگری از اهداف موجود پیدا کنند^۱.

۲. برای داروهای جدید که طراحی شده و پروسه‌ی سخت تایید را پشت سر می‌گذارند، اهداف دیگری از اهداف جدید معرفی کنند^۲.

۳. برای اهداف جدیدی که شناسایی می‌شوند از بین داروهای موجود داروی مناسبی برای آن پیدا کنند^۲.

اخیرا شرکت‌های بزرگ داروسازی دنیا نظیر شرکت‌های فایزر آمریکا، نواریتز سوئیس و بایر آلمان بخش‌های تحقیقاتی برای این منظور راه‌اندازی کرده‌اند و آینده‌ی علم داروسازی را در این دیدگاه جدید می‌دانند. اولین و مهم‌ترین مرحله از دیدگاه فوق، شناسایی یا به طور دقیق‌تر پیش‌بینی تعامل‌های مخفی بین داروها و اهداف (پروتئین‌ها) است. روش‌های تجربی برای این کار پرهزینه و زمان‌بر هستند و فقط زمانی می‌توانند استفاده شوند که تعداد هدف‌ها و داروها محدود باشند. درمقابل روش‌های محاسباتی توانایی کار با حجم عظیمی از داروها و هدف‌ها را دارند. این کار معمولاً با استفاده از اطلاعات تعامل‌هایی که تاکنون شناسایی شده‌اند، اطلاعات میزان شباهت داروها به هم و اطلاعات میزان شباهت اهداف به هم انجام می‌پذیرد. مجموعه داده‌هایی وجود دارند که این اطلاعات را شامل می‌شوند. تاکنون چندین الگوریتم برای پیش‌بینی تعامل بین دارو و هدف معرفی شده‌اند که با رویکرد خاص خود سعی در ارائه‌ی پیش‌بینی خود داده‌اند. از جمله‌ی آن‌ها روش‌های آماری، روش‌های یادگیری ماشین، روش‌های معادلات دیفرانسیل، روش‌های مبتنی بر گراف و ... هستند چنگ و گولیمین [۱۰]، هی و ژانگ [۷]، پرلمن و گاتلیب [۶]، شن و لیو [۴] می و وه [۲].

در این بین می‌توان به روش مدل تعامل گراف در [۸] اشاره کرد. یک روش استنتاج همراه با ناظر است که براساس یک گراف دویخشی جهت پیش‌بینی تعامل‌های ناشناخته‌ی هدف و داروهای استفاده می‌شود. در [۵] از روش استنتاج براساس شبکه‌ی (NBI) استفاده شده که نسخه‌ای ساده شده از الگوریتم ارائه شده در [۹] را پیاده‌سازی کرده است. این روش نتایج کارایی بالایی را نشان می‌داد اما از شباهت داروها و شباهت اهداف (پروتئین‌ها) استفاده نشده است. محدودیت دیگر NBI ناتوانی این الگوریتم در ارائه‌ی پیش‌بینی برای داروهای جدید که هدف تعیین شده‌ی قبلی ندارند می‌باشد. محدودیت اول NBI در روش DT-Hybrid ارائه شده در [۴] برطرف شد این روش اطلاعات مربوط به شباهت داروها و شباهت پروتئین‌ها را در مدل خود در نظر گرفت. روش موسوم به HGBI در [۳] اطلاعات مربوط به شباهت داروها و شباهت پروتئین‌ها را در قالب دو گراف شباهت به مدل اضافه کرد. این روش دو محدودیت NBI را از پیش رو برداشت و نتایج بهتری به دست آورد.

این مسئله در سال‌های اخیر با سرعت بیشتری مورد توجه قرار گرفته است که طبق اطلاع نگارنده فقط در سال ۲۰۱۷ بیش از ۱۵ مقاله و روش جدید در این مبحث ارائه و چاپ شده است. یعنی نه تنها مسئله قدیمی نشده و برای همیشه حل نشده است بلکه رفته رفته پخته‌تر شده و روش‌های جدیدتر و پیچیده‌تر در آن به کار رفته است.

وقتی تعداد مقالات در یک حوزه به این شکل افزایش می‌یابد، نیاز به مقالات مروری افزایش می‌یابد. مقالاتی که روش‌های مطرح را معرفی و مقایسه کنند و معیارهای مناسبی را جهت مقایسه‌ی رویکردهای مختلف حل مسئله ارائه دهند. مقالات مروری موجود به تعداد اندکی موجود بوده و در این زمینه فقط به معرفی روش‌ها و رویکردهای روش‌ها و دسته‌بندی این روش‌ها در گروه‌هایی براساس شباهت بین آن‌ها پرداخته است. تعداد مقالات این حوزه زیاد است و هر مقاله براساس روش‌های خاص نویسندگانش با متدهای دیگر مقایسه شده است و از روال خاصی پیروی نکرده‌اند. این خود باعث ایجاد ناعدالتی در مقایسه‌ها می‌شود و گاهی نتایجی ارائه می‌شود که عمدی یا سهوی دستکاری شده‌اند. در این شرایط احساس نیاز به مقالات گزارشی که بسته‌ای جهت مقایسه ارائه دهند وجود دارد.

۳-۶- روش انجام تحقیق و مراحل اجراء (در صورت امکان جدول زمان‌بندی ارائه شود):

در این کار برآنیم مقالات و کارهای قبلی انجام شده در این زمینه را مطالعه و بررسی شود. چند نرم‌افزار الگوریتم‌های معروف را در صورت متن‌باز بودن دریافت و درغیراینصورت پیاده‌سازی خواهد شد. سپس رویکرد الگوریتم داری را در قالب یک بسته

² New drug

³ New target

⁴ network-based inference

ارائه شود تا نرم افزار الگوریتم های در این زمینه را باهم مقایسه کند. جهت ایجاد مقایسه ای عادلانه از داده های استاندارد طلایی استفاده خواهد شد.

۴ مجموعه داده ی طلایی استاندارد برای تحقیقات DTI وجود دارد که اکثر کارها بر روی این ۴ مجموعه داده انجام شده اند. این مجموعه داده ها شامل nuclear و G-protein-coupled receptors (GPCRs) ، ion channels. human enzymes receptors می باشند که در [۹] جمع آوری شده است. این مجموعه داده ها، اطلاعات مفصلی در رابطه با داروهایی که تا کنون طراحی شده اند، ساختار شیمیایی آن ها، تعامل هایی که با هدف های دارویی دارند و ساختار توالی هدف های دارویی را شامل می شود. از این داده ها برای ارزیابی همه ی روش هایی که برای پیش بینی تعامل بین دارو و هدف ارائه شده اند، استفاده خواهد شد.

	Nuclear Receptor	GPCR	Ion Channel	Enzyme
Number of drugs	54	223	210	445
Number of targets	26	95	204	664
Number of interaction pairs	90	635	1476	2926
Average number of drugs per target	3.46	6.68	7.24	4.41
Average number of targets per drug	1.67	2.85	7.03	6.58
Sparsity of the interaction matrix	93.59%	97.00%	96.55%	99.01%
Percentage of drugs that have only one interaction target	72.22%	47.53%	38.57%	39.78%
Percentage of targets that have only one interaction drug	30.77%	35.79%	11.27%	43.37%

در این راه داده های اصلی مورد استفاده جهت آموزش و ارزیابی ۴ مجموعه داده ی طلایی بیان شده در بالا خواهند بود.

روش های ارزیابی K-Fold Cross Validation خواهد بود به این ترتیب داده را به k بخش تقسیم کرده و یک قسمت را به عنوان داده ی اعتبارسنجی و بقیه را به عنوان داده ی آموزشی در نظر گرفته می شود و این کار برای k-1 بخش دیگر نیز انجام می شود. معیارهای ارزیابی این روش ها AUC و AUPR خواهد بود.

ارزیابی:

برای سادگی معیارهای ارزیابی الگوریتم های دسته بندی، آنها را برای یک مسئله با دو دسته ارائه خواهیم نمود. در ابتدا با مفهوم ماتریس درهم ریختگی (Classification Matrix) آشنا می شویم. این ماتریس چگونگی عملکرد الگوریتم دسته بندی را با توجه به مجموعه داده ورودی به تفکیک انواع دسته های مساله دسته بندی، نمایش می دهد.

		Predicted	
		Positive	Negative
Actual	Positive	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

هر یک از عناصر ماتریس به شرح ذیل می باشد :

TN: بیانگر تعداد رکوردهایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته بندی نیز دسته آنها را بدرستی منفی تشخیص داده است.

TP: بیانگر تعداد رکوردهایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته بندی نیز دسته آنها را بدرستی مثبت تشخیص داده است.

FP: بیانگر تعداد رکوردهایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته بندی دسته آنها را به اشتباه مثبت تشخیص داده است.

FN: بیانگر تعداد رکوردهایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته بندی دسته آنها را به اشتباه منفی تشخیص داده است.

مهمترین معیار برای تعیین کارایی یک الگوریتم دسته بندی دقت یا نرخ دسته بندی^۵ (CA) است که این معیار دقت کل یک دسته بند را محاسبه می کند. در واقع این معیار مشهورترین و عمومی ترین معیار محاسبه کارایی الگوریتم های دسته بندی است که نشان می دهد، دسته بند طراحی شده چند درصد از کل مجموعه رکوردهای آزمایشی را بدرستی دسته بندی کرده است . دقت دسته بندی با استفاده از رابطه I بدست می آید که بیان می کند دو مقدار TP و TN مهمترین مقادیری هستند که در یک مسئله دودسته ای باید بیشینه شوند (در مسائل چند دسته ای مقادیر قرار گرفته روی قطر اصلی این ماتریس - که در صورت کسر محاسبه CA قرار می گیرند - باید بیشینه باشند).

معیار خطای دسته بندی^۶ (ER) دقیقاً برعکس معیار دقت دسته بندی است که با استفاده از رابطه II بدست می آید. کمترین مقدار آن برابر صفر است زمانی که بهترین کارایی را داریم و بطور مشابه بیشترین مقدار آن برابر یک است زمانی که کمترین کارایی را داریم.

ذکر این نکته ضروری است که در مسائل واقعی، معیار دقت دسته بندی به هیچ عنوان معیار مناسبی برای ارزیابی کارایی الگوریتم های دسته بندی نمی باشد، به این دلیل که در رابطه دقت دسته بندی، ارزش رکوردهای دسته های مختلف یکسان در نظر گرفته می شوند. بنابراین در مسائلی که با دسته های نامتعادل سروکار داریم، به بیان دیگر در مسائلی که ارزش دسته ای در مقایسه با دسته دیگر متفاوت است، از معیارهای دیگری استفاده می شود .

همچنین در مسائل واقعی معیارهای دیگری نظیر DR و FAR که به ترتیب از روابط III و IV بدست می آیند، اهمیت ویژه ای دارند. این معیارها که توجه بیشتری به دسته بند مثبت نشان می دهند، توانایی دسته بند را در تشخیص دسته مثبت و بطور مشابه تاوان این توانایی تشخیص را تبیین می کنند. معیار DR نشان می دهد که دقت تشخیص دسته مثبت چه مقدار است و معیار FAR نرخ هشدار غلط را با توجه به دسته منفی بیان می کند.

$$I: \quad CA = \frac{TN+TP}{TN+FN+TP+FP}$$

$$II: \quad ER = \frac{FN+FP}{TN+FN+TP+FP} = 1 - CA$$

$$III: \quad DR = \frac{TP}{FN+TP}$$

$$IV: \quad FAR = \frac{FP}{TN+FP}$$

$$V: \quad \text{Precision} = TP / TP + FP$$

$$VI: \quad \text{Recall} = TP / TP + FN$$

معیار مهم دیگری که برای تعیین میزان کارایی یک دسته بند استفاده می شود معیار^۷ AUC است.

AUC نشان دهنده سطح زیر نمودار ROC^۸ می باشد که هر چه مقدار این عدد مربوط به یک دسته بند بزرگتر باشد

کارایی نهایی دسته بند مطلوب تر ارزیابی می شود. نمودار ROC روشی برای بررسی کارایی دسته بندها می باشد. در واقع منحنی های ROC منحنی های دو بعدی هستند که در آنها DR یا همان نرخ تشخیص صحیح دسته مثبت^۹ (TPR) روی محور Y و بطور مشابه

5 Classification Accuracy - Rate

6 Error Rate

7 Area Under Curve

8 Receiver Operating Characteristic

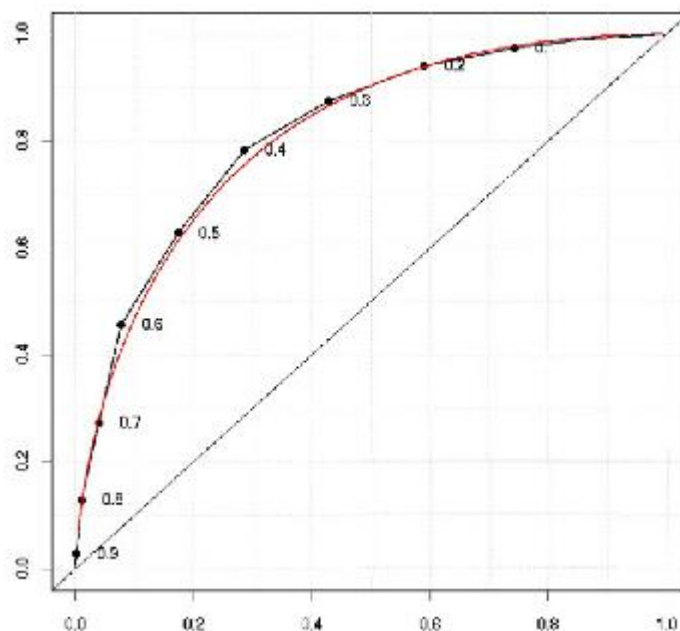
9 True Positive Rate

FAR یا همان نرخ تشخیص غلط دسته منفی ۱۰ (FPR) روی محور X رسم می‌شوند. به بیان دیگر یک منحنی ROC مصالحه نسبی میان سودها و هزینه‌ها را نشان می‌دهد.

بسیاری از دسته بندها همانند روش‌های مبتنی بر درخت تصمیم و یا روش‌های مبتنی بر قانون، به گونه ای طراحی شده اند که تنها یک خروجی دودویی (مبنی بر تعلق ورودی به یکی از دو دسته ممکن) تولید می‌کنند. به این نوع دسته بندها که تنها یک خروجی مشخص برای هر ورودی تولید می‌کنند، دسته بندهای گسسته گفته می‌شود که این دسته بندها تنها یک نقطه در فضای ROC تولید می‌کنند.

بطور مشابه دسته بندهای دیگری نظیر دسته بندهای مبتنی بر روش بیز و یا شبکه‌های عصبی نیز وجود دارند که یک احتمال و یا امتیاز برای هر ورودی تولید می‌کنند، که این عدد بیانگر درجه تعلق ورودی به یکی از دو دسته موجود می‌باشد. این دسته بندها پیوسته نامیده می‌شوند و بدلیل خروجی خاص این دسته بندها یک آستانه جهت تعیین خروجی نهایی در نظر گرفته می‌شود.

یک منحنی ROC اجازه مقایسه تصویری مجموعه ای از دسته بندی کننده‌ها را می‌دهد، همچنین نقاط متعددی در فضای ROC قابل توجه است. نقطه پایین سمت چپ (0,0) استراتژی را نشان می‌دهد که در یک دسته بند مثبت تولید نمی‌شود. استراتژی مخالف، که بدون شرط دسته بندهای مثبت تولید می‌کند، با نقطه بالا سمت راست (1,1) مشخص می‌شود. نقطه (0,1) دسته بندی کامل و بی عیب را نمایش می‌دهد. بطور کلی یک نقطه در فضای ROC بهتر از دیگری است اگر در شمال غربی تر این فضا قرار گرفته باشد. همچنین در نظر داشته باشید منحنی‌های ROC رفتار یک دسته بندی کننده را بدون توجه به توزیع دسته‌ها یا هزینه خطا نشان می‌دهند، بنابراین کارایی دسته بندی را از این عوامل جدا می‌کنند. فقط زمانی که یک دسته بند در کل فضای کارایی به وضوح بر دسته دیگری تسلط یابد، می‌توان گفت که بهتر از دیگری است. به همین دلیل معیار AUC که سطح زیر نمودار ROC را نشان می‌دهد می‌تواند نقش تعیین کننده ای در معرفی دسته بند برتر ایفا کند. برای درک بهتر نمودار ROC زیر را مشاهده کنید.



مقدار AUC برای یک دسته بند که بطور تصادفی، دسته نمونه مورد بررسی را تعیین می‌کند برابر 0.5 است. همچنین بیشترین مقدار این معیار برابر یک بوده و برای وضعیتی رخ می‌دهد که دسته بند ایده آل بوده و بتواند کلیه نمونه‌های مثبت را بدون هرگونه هشدار غلطی تشخیص دهد.

معیار AUPR در نمودار PR بر روی محور Yها به جای TPR مقادیر Precision قرار می‌گیرد و به جای FPR بر روی محور Xها Recall که همان TPR است قرار می‌گیرد. به مساحت زیر این نمودار ROC، AUPR گویند. این معیار همانند معیار

AUC محاسبه شده و در مقایسه‌ها به کار می‌رود. با این تفاوت که در هنگام مقایسه اگر دسته‌ها در اصل نامتعادل باشند به طور مثال در یک دو دسته‌ای تعداد دسته‌ی مثبت خیلی کمتر از تعداد دسته‌ی منفی باشد آنگاه معیار AUPR بهتر از AUC می‌تواند توانایی دسته‌بندی را برای تشخیص هر چه بهتر گروه مثبت مقایسه کند.

معیار AUC و AUPR برخلاف دیگر معیارهای تعیین کارایی دسته‌بندیها مستقل از آستانه تصمیم‌گیری دسته‌بندی می‌باشند. بنابراین این معیارها نشان دهنده میزان قابل اعتماد بودن خروجی یک دسته‌بندی مشخص به ازای مجموعه داده‌های متفاوت هستند که این مفهوم توسط سایر معیارهای ارزیابی کارایی دسته‌بندیها قابل محاسبه نمی‌باشد.

حالت‌های آزمون k-fold cross validation:

ماتریس تعامل‌های شناخته‌شده بین دارو و هدف را در نظر بگیرید به صورتی که در ستون‌ها اهداف و در ردیف‌ها داروها قرار گیرند آنگاه برای آزمودن الگوریتم‌ها از سه حالت زیر استفاده می‌شود:

۱- حذف یک درایه از ماتریس یال‌های بین دارو و هدف و اجرای الگوریتم جهت برگرداندن یال حذف‌شده (شبیه‌سازی Drug repositioning).

۲- حذف یک ستون و اجرای الگوریتم جهت برگرداندن آن ستون. در این حالت تمام روابط یک هدف با بقیه‌ی داروها حذف می‌شود. (شبیه‌سازی New Drug).

۳- حذف یک ردیف و اجرای الگوریتم جهت برگرداندن آن ردیف. در این حالت تمام روابط یک دارو با بقیه‌ی اهداف حذف می‌شود. (شبیه‌سازی New Target).

(A) known drug - known target					(B) known drug - new target				
	T1	T2	T3	T4		T1	T2	T3	T4
D1	0	1	0	1	D1	0	1	0	1
D2	1				D2	0			
D3	0				D3	0			
D4	1				D4	0			
D5	0				D5	0			

(C) new drug - known target				
	T1	T2	T3	T4
D1	0	0	0	0
D2	1			
D3	0			
D4	1			
D5	0			

- [3] Wang W, Yang S, Li J (2013) Drug target predictions based on heterogeneous graph inference. *Pac Symp Biocomput* 18:53–64
- [4] Chen, X., Liu, M. X. and Yan, G. Y. (2012), Drug–target interaction prediction by random walk on the heterogeneous network., *Molecular BioSystems* 8, 7, 1970-1978.
- [5] Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, Zhou W, Huang J, Tang Y (2012) Prediction of drug–target interactions and drug repositioning via network-based inference. *PLoS Comput Biol* 8(5):e1002503
- [6] Perlman, L., Gottlieb, A., Atias, N., Ruppin, E. and Sharan, R. (2011), Combining drug and gene similarity measures for drug-target elucidation., *Journal of computational biology* 18, 2, 133-145.
- [7] He, Z., Zhang, J., Shi, X. H., Hu, L. L., Kong, X., Cai, Y. D. and Chou, K. C. (2010), Predicting drug-target interaction networks based on functional groups and biological features., *PloS one* 5, 3, e9603.
- [8] Bleakley K, Yamanishi Y (2009) Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 25(18):2397–2403
- [9] Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M (2008) Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24(13):i232–i240
- [10] Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* 25(1):71–75
- [11] Zhou T, Ren J, Medo M, Zhang YC (2007) Bipartite network projection and personal recommendation. *Phys Rev E Stat Nonlinear Soft Matter Phys* 76(4 Pt 2):04611

۳-۸- نهاد یا مؤسساتی که می‌توانند از یافته‌های این پژوهش بهره گیرند:

۴- مشخصات استاد (استادان) مشاور (پیشنهادی):

استاد مشاور اول	
* نام و نام خانوادگی:	* رشته – گرایش:
محل خدمت:	رتبه دانشگاهی:
سال و محل اخذ آخرین مدرک تحصیلی:	آدرس پست الکترونیکی:
سنوات تدریس در دوره کارشناسی ارشد:	تعداد پایان نامه‌های کارشناسی ارشد راهنمایی شده:
* تعداد پایان نامه‌های کارشناسی ارشد در دست راهنمایی:	
استاد مشاور دوم	
* نام و نام خانوادگی:	* رشته – گرایش:
محل خدمت:	رتبه دانشگاهی:
سال و محل اخذ آخرین مدرک تحصیلی:	آدرس پست الکترونیکی:
سنوات تدریس در دوره کارشناسی ارشد:	تعداد پایان نامه‌های کارشناسی ارشد راهنمایی شده:
* تعداد پایان نامه‌های کارشناسی ارشد در دست راهنمایی:	

* برای اعضای هیات علمی دانشگاه شهید بهشتی، تکمیل تنها قسمت‌های ستاره‌دار لازم است.

تذکر 1: نتایج تحقیقات و پایان نامه‌های دانشجویان و درآمدهای ناشی از آن متعلق به دانشگاه شهید بهشتی بوده و هیچ شخصیت حقیقی و حقوقی بدون اجازه دانشگاه حق بهره‌برداری از آن را نخواهد داشت. همچنین کلیه حقوق پایان نامه اعم از چاپ، تکثیر، نسخه برداری، ترجمه، اقتباس، و نظایر آن برای دانشگاه شهید بهشتی محفوظ است. نقل مطالب با ذکر مأخذ بلامانع است.

اظہارنامہ دانشجو:

بدین وسیلہ تأکید می‌شود بر اساس نتایج جستجو، موضوع پایان نامہ تکراری نیست.

نام و نام خانوادگی دانشجو:

محمد امین خدامرادی امضاء: تاریخ: ۱۳۹۶/۰۸/۱۶

نام و نام خانوادگی استاد (استادان) راهنما:

۱- دکتر چنگیز اصلاحچی امضاء: تاریخ: ۱۳۹۶/۰۸/۱۶

۲- امضاء: تاریخ: