Data Complexity measures in feature selection

روند کلی الگوریتم:

- ویژگیها توسط univariate complexity measure ها رتبه بندی شوند
این فرآیند با هزینه کمی انجام میشود و کمک میکند که ویژگیهای غیرمرتبط حذف شوند
- ویژگیهایی که به کمک مرحله قبل در رتبه های بالا قرار گرفتند توسط دیگر معیارها بصورت multivariante بررسی شوند و ویژگیهای مرتبط انتخاب شوند

Univariate FS

در این مرحله از معیارهای F1, F2, F3 , F4 استفاده میشود
برای رتبه بندی ویژگیها از سه روش استفاده شده است:

1- The first computes the precision of feature ranking when a threshold corresponding to the known number of relevant features is used to select the features.
2- The second evaluation metric computes the percentage of the ranking (Coverage %) that has to be regarded in order to retrieve all relevant features.
3- The third metric is based on the AUC (Area Under the ROC curve) concept, which is independent of a particular threshold value on the number of chosen features.

بعد از بررسی رفتار این معیارها به کمک دیتاستهای گوناگون نتیجه شده است که:
معیار F1 بهترین نتایج را دارد معیار F2 بهترین precision را کسب میکند و معیار F3 converage بهتری دارد و نهایتا معیار F4 ضعیفترین عملکرد را دارد

Multivariate FS

ورودی این مرحله خروجی مرحله قبلی است. برای انتخاب فیچرها در مرحله قبلی باید تا جایی به حذف ویژگیها ادامه دهیم که نهایتا بتوانیم به ۹۵ درصد از صحت استفاده از همه ی ویژگی ها برسیم

برای این مرحله از دو دیدگاه استفاده شده است: Forward and backward selection
برای اینکه بررسی کنیم کدام یک از این دو دیدگاه موثرتر است دو معیار N1 , N2 را با هر دو حالت محاسبه میکنیم
معیار N2 در زمان forward selection نتایج بهتری دارد اما N1 در هر دو جهت خوب عملکرده و نتایج حدودا یکسانی بدست می آید
بصورت کلی میتوان نتیجه گرفت که forward selection خروجی بهتری دارد
نتایها در مرحله دوم از روابط زیر برای انتخاب ویژگیهای نهایی استفاده میکنیم:

- if $m < 10$, select 75% of the features;
- if $10 \leq m < 75$, select 40% of the features;
- if $75 \leq m < 100$, select 10% of the features;
- if $m \geq 100$, select 3% of the features.

Figure 4. Univariate-multivariate feature selection (UMFS) algorithm framework.