frlily / October 11, 2015 12:56PM

資料分析領域中最為人稱道的七種降維方法

近來由於資料記錄和屬性規模的急劇增長，巨量資料處理平台和並行[url=http://www.finereport.com/tw/]資料分析[/url]演算法也隨之出現。於此同時，這也推動了資料降維處理的應用。實際上，資料量有時過猶不及。有時在資料分析應用中大量的資料反而會產生更壞的性能。

最新的一個例子是採用 2009 KDD Challenge 巨量資料集來預測客戶流失量。 該資料集維度達到 15000 維。 大多數資料挖掘演算法都直接對資料逐列處理，在資料數目一大時，導致演算法越來越慢。該項目的最重要的就是在減少資料列數的同時保證丟失的資料信息儘可能少。

以該項目為例，我們開始來探討在當前資料分析領域中最為資料分析人員稱道和接受的資料降維方法。

缺失值比率 (Missing Values Ratio)
該方法的是基於包含太多缺失值的資料列包含有用信息的可能性較少。因此，可以將資料列缺失值大於某個閾值的列去掉。閾值越高，降維方法更為積極，即降維越少。該方法示意圖如下：

[img]http://www.finereport.com/tw/wp-content/themes/BusinessNews/images/15101301.bmp[/img]
低方差濾波 (Low Variance Filter)
與上個方法相似，該方法假設資料列變化非常小的列包含的信息量少。因此，所有的資料列方差小的列被移除。需要注意的一點是：方差與資料範圍相關的，因此在採用該方法前需要對資料做歸一化處理。演算法示意圖如下：

[img]http://www.finereport.com/tw/wp-content/themes/BusinessNews/images/15101302.bmp[/img]
高相關濾波 (High Correlation Filter)
高相關濾波認為當兩列資料變化趨勢相似時，它們包含的信息也顯示。這樣，使用相似列中的一列就可以滿足機器學習模型。對於數值列之間的相似性通過計算相關係數來表示，對於名詞類列的相關係數可以通過計算皮爾遜卡方值來表示。相關係數大於某個閾值的兩列只保留一列。同樣要注意的是：相關係數對範圍敏感，所以在計算之前也需要對資料進行歸一化處理。演算法示意圖如下：

[img]http://www.finereport.com/tw/wp-content/themes/BusinessNews/images/15101303.bmp[/img]
隨機森林/組合樹 (Random Forests)
組合決策樹通常又被成為隨機森林，它在進行特徵選擇與構建有效的分類器時非常有用。一種常用的降維方法是對目標屬性產生許多巨大的樹，然後根據對每個屬性的統計結果找到信息量最大的特徵子集。例如，我們能夠對一個非常巨大的資料集生成非常層次非常淺的樹，每顆樹只訓練一小部分屬性。如果一個屬性經常成為最佳分裂屬性，那麼它很有可能是需要保留的信息特徵。對隨機森林資料屬性的統計評分會向我們揭示與其它屬性相比，哪個屬性才是預測能力最好的屬性。演算法示意圖如下：

[img]http://www.finereport.com/tw/wp-content/themes/BusinessNews/images/15101304.bmp[/img]
主成分分析 (PCA)
主成分分析是一個統計過程，該過程通過正交變換將原始的 n 維資料集變換到一個新的被稱做主成分的資料集中。變換後的結果中，第一個主成分具有最大的方差值，每個後續的成分在與前述主成分正交條件限制下與具有最大方差。降維時僅保存前 m(m PCA

[img]http://www.finereport.com/tw/wp-content/themes/BusinessNews/images/15101305.bmp[/img]
反向特徵消除 (Backward Feature Elimination)
在該方法中，所有分類演算法先用 n 個特徵進行訓練。每次降維操作，採用 n-1 個特徵對分類器訓練 n
次，得到新的 n 個分類器。將新分類器中錯分率變化最小的分類器所用的 n-1
維特徵作為降維後的特徵集。不斷的對該過程進行迭代，即可得到降維後的結果。第k 次迭代過程中得到的是 n-k 維特徵分類器。通過選擇最大的錯誤容忍率，我們可以得到在選擇分類器上達到指定分類性能最小需要多少個特徵。演算法示意圖如下：
Backward Feature Elimination

[img]http://www.finereport.com/tw/wp-content/themes/BusinessNews/images/15101306.bmp[/img]
前向特徵構造 (Forward Feature Construction)
前向特徵構建是反向特徵消除的反過程。在前向特徵過程中，我們從 1 個特徵開始，每次訓練添加一個讓分類器性能提升最大的特徵。前向特徵構造和反向特徵消除都十分耗時。它們通常用於輸入維數已經相對較低的資料集。演算

法示意圖如下：

[img]http://www.finereport.com/tw/wp-content/themes/BusinessNews/images/15101307.bmp[/img]
我們選擇 2009 KDD chanllenge 的削資料集來對這些降維技術在降維率、準確度損失率以及計算速度方面進行比較。當然，最後的準確度與損失率也與選擇的資料分析模型有關。因此，最後的降維率與準確度的比較是在三種模型中進行，這三種模型分別是：決策樹，神經網路與樸素貝葉斯。

通過運行優化循環，最佳循環終止意味著低緯度與高準確率取決於七大降維方法與最佳分類模型。最後的最佳模型的性能通過採用所有特徵進行訓練模型的基準準確度與 ROC
曲線下的面積來進行比較。下面是對所有比較結果的對比。

[img]http://www.finereport.com/tw/wp-content/themes/BusinessNews/images/15101308.png[/img]

從上表中的對比可知，資料降維演算法不僅僅是能夠提高演算法執行的速度，同時也能過提高分析模型的性能。
在對資料集採用：缺失值降維、低方差濾波，高相關濾波或者隨機森林降維時，表中的 AoC
在測試資料集上有小幅度的增長。

[img]http://www.finereport.com/tw/wp-content/themes/BusinessNews/images/15101309.bmp[/img]
確實在巨量資料時代，資料越多越好似乎已經成為公理。我們再次解釋了當資料資料集寶航過多的資料雜訊時，演算法的性能會導致演算法的性能達不到預期。移除信息量較少甚至無效信息唯獨可能會幫助我們構建更具擴展性、通用性的資料模型。該資料模型在新資料集上的表現可能會更好。

最近，我們諮詢了 LinkedIn 的一個資料分析小組在資料分析中最為常用的資料降維方法，除了本博客中提到的其中，還包括：隨機投影(Random Projections)、非負矩陣分解(N0n-negative Matrix
Factorization),自動編碼(Auto-encoders),卡方檢測與信息增益(Chi-square and information gain)，
多維標定(Multidimensional Scaling), 相關性分析(Coorespondence Analysis), 因子分析(Factor
Analysis)、聚類(Clustering)以及貝葉斯模型(Bayesian Models)。

分享自： 数盟社区