大家好，今天我來匯報的論文內容是《利用支援向量機在更小的訓練集合裏訓練反垃圾郵件模型》。

衆所周知，現在是在一個互聯網的時代，是一個信息的時代，隨著科學技術的蓬勃發展，人們的生活發生了很大的變化。

舉個例子來説，Email 的誕生，就是一個影響重大的事件，就在現在，上百萬上千萬的電子郵件正在被發送或者接受。但是這裏面究竟有多少是有價值的郵件呢？

科學技術帶來方便性的同時，也帶來了一些麻煩，仿佛是一把雙刃劍一樣。科學家們也正在研究如何去封鎖垃圾郵件。

由於真實世界中的垃圾郵件的數量實在是太過於龐大了，而且隨著垃圾郵件越來越狡猾，簡單基於黑名單的技術已經無法滿足日常的使用了，我們需要一個更先進的更智能的能夠自主學習的垃圾郵件過濾系統。

解決這個難題最新的進展是使用機器學習的方法，典型的來説，機器學習方法把垃圾郵件過濾問題轉換成一個二元分類任務（binary classification task）。他首先從標記過的訓練數據裏面學習分類器，然後使用分類器對接下來傳入的數據進行分類。在所有候選的機器學習方法中，SVM 由於它獨特的特性，深受企業用戶的好評。

雖然機器學習方法在是實驗室裏面有著非常好的性能，但是在企業實際情況的應用中遠遠不能滿足要求。

大量垃圾郵件通過自動運行的程式發出，已經到達一個驚人的數量級。由於普通郵件個人隱私的考量，普通郵件和垃圾郵件之間有很大的分發差距。在企業環境中，你需要頻繁升級垃圾郵件分類器才能保證日常使用不會被干擾。

總而言之，問題主要是如何保同時保證效率和性能。雖然已經提出了各種方法，但是目前這些解決方案并不符合我們現在實際要求的性能和效率上的要求。

本篇論文提出了一種新方法。采用最小信息損失的壓縮方案去壓縮訓練集合。這個方法關鍵的過程是需要減少冗餘，舉例來説，目前在開發的一個軟件，可以清洗樣本集，刪除重複或者非常相似的樣本。這樣做可以大量減少訓練過程中需要采樣的郵件數量，從而可以提升結果的精確度和及時性。經過實驗研究表明，使用這個方法的訓練子集比使用原始數據有著更好的性能。

我在這邊不在贅述關於支援向量機相關的技術要點，相信大家已經對 SVM 解決問題的方法有一定的瞭解。

以前解決大量訓練集工作的方法通常表現為使用不同的 SVM 實現技術，例如 SMO（序列最小優化算法-由微軟提出），目前廣汎在 SVM 的訓練過程中使用。

回到原來的話題，我們需要解決垃圾郵件的問題，我們需要高性能且有效的方法處理一個大型的訓練數據集合。我們通過以下步驟來進行訓練數據集合的處理。見圖。
爲了更好的解決冗餘問題，我們對正數據也執行采樣，雖然相對較小，但是絕對大小也很大，并且不需要保留全部。
一般來説，這個解決方法可以歸納為：
· 使用 SV 作为减少训练池的基础，以保证分类的精度。
·使用冗余减少工具根据 SV 计算的垃圾邮件发送率以减少培训电子邮件样本集。
· 组合 SV 电子邮件和减少来自原始培训池的电子邮件样本，以在删除重复项之后建立最终平衡的壓縮過的樣本集合。

針對這個解決方案進行了實驗，實驗使用的訓練數據集合大小是 64880，測試數據集合是 280 904，訓練數據集合裏面包含了 5416 份正常的郵件和 59464 份垃圾郵件。訓練數據集和測試數據集之間并沒有重叠。

在機器學習中有四个基本值来反映分类器的性能，分别是 FALSE POSITIVE,TRUE POSITIVE,FALSE NEGATIVE, TRUE NEGATIVE. Acc+指的是 positive 的 data，acc-指的是 negative 的 data。此外，我们使用加权几何平均值而不是整体精度，以评估我们的方法的整体性能。

實驗使用了不同的壓縮率進行了六次實驗來壓縮原始數據集合。

實驗結果表明實現對訓練集合進行壓縮可以顯著減少訓練集合的大小。這可以帶來效率和性能上的提升，實驗同時表明，使用壓縮過的集合進行訓練和使用原始集合進行訓練，差距是很小的。但是帶來的提升卻很巨大。

雖然實驗是成功的，但是要將這個解決方案帶進實際環境中還有很多工作需要做，不過這也為解決性能和效率問題提供了全新的思路。