phức tạp của giải thuật máy học SVM chuẩn về giải hệ phương trình tuyến tính đơn giản hơn. Nghiên cứu của (Liu *et al*., 1999), (Poulet & Do, 2004) đã đề nghị xây dựng giải thuật học tăng trưởng, chỉ nạp dữ liệu từng phần rồi cập nhật mô hình theo dữ liệu mà không cần nạp toàn bộ tập dữ liệu trong bộ nhớ. Công trình nghiên cứu của (Do & Poulet, 2004), (Do & Poulet, 2006), (Do & Poulet, 2008) đề nghị giải thuật song song để cải thiện tốc độ huấn luyện. (Tong & Koller, 2000), (Do & Poulet, 2005) đề nghị phương pháp chọn tập con dữ liệu thay vì phải học trên toàn bộ tập dữ liệu gốc. (Do & Fekete, 2007) kết hợp boosting (Freund & Schapire, 1999), arcing (Breiman, 1997) để cải thiện tốc độ xây dựng mô hình SVM chỉ tập trung vào những mẫu khó phân lớp.

Nghiên cứu của chúng tôi trong bài viết này nhằm phát triển từ ý tưởng sử dụng giải thuật giảm gradient ngẫu nhiên (SGD) để giải trực tiếp vấn đề tối ưu của máy học SVM, được đề xuất bởi (Bottou & Bousquet, 2008) và (Shalev-Shwartz *et al*., 2007). Tuy nhiên, vấn đề tối ưu của máy học SVM có hàm hinge loss không khả vi là nguyên nhân ảnh hưởng đến hiệu quả của giải thuật SGD. Chúng tôi đề xuất thay thế hàm hinge loss bằng các hàm xấp xỉ, khả vi (Rennie, 2004) để cải tiến tốc độ hội tụ của giải thuật SGD. Kết quả thực nghiệm trên 2 tập dữ liệu văn bản lớn RCV1 (Bottou & Bousquet, 2008), twitter (Go *et al*., 2009) cho thấy hiệu quả của đề xuất sử dụng hàm xấp xỉ so với hàm hinge loss.

Phần tiếp theo của bài được tổ chức như sau. Phần 2 sẽ trình bày tóm tắt về máy học SVM, giải thuật SGD sử dụng trong SVM và thay thế hàm hinge loss bằng các hàm xấp xỉ khả vi, cải tiến tốc độ hội tụ của giải thuật SGD. Kết quả chạy thử nghiệm sẽ được trình bày trong phần 3 trước khi kết thúc bằng kết luận và hướng phát triển.

2 GIẢI THUẬT GIẢM GRADIENT NGẪU NHIÊN CHO VẤN ĐỀ PHÂN LỚP CỦA MÁY HỌC VÉC-TƠ HỖ TRỢ

2.1 Phân lớp với máy học véc-tơ hỗ trợ

Xét ví dụ phân lớp nhị phân tuyến tính như Hình 1. Cho m phần tử $x_1, x_2, \ldots, x_m$ trong không gian n chiều với nhãn (lớp) của các phần tử tương ứng là $y_1, y_2, \ldots, y_m$ có giá trị 1 hoặc -1. Nhãn $y_i = 1$ khi x_i thuộc lớp $+1$ (lớp dương, lớp chúng ta quan tâm) và $y_i = -1$, nếu x_i thuộc lớp -1 (lớp âm hay các lớp còn lại). SVM tìm siêu phẳng tối ưu (xác định bởi véc tơ pháp tuyến w và độ lệch của siêu phẳng b) dựa trên 2 siêu phẳng hỗ trợ của 2 lớp. Các phần tử lớp $+1$ nằm bên phải của siêu phẳng hỗ trợ cho lớp $+1$, các phần tử lớp -1 nằm phía bên trái của siêu phẳng hỗ trợ cho lớp -1. Những phần tử nằm ngược phía với siêu phẳng hỗ trợ được coi như lỗi. Khoảng cách lỗi được biểu diễn bởi $z_i \geq 0$ (với x_i nằm đúng phía của siêu phẳng hỗ trợ của nó thì khoảng cách lỗi tương ứng $z_i = 0$, còn ngược lại thì $z_i > 0$ là khoảng cách từ điểm x_i đến siêu phẳng hỗ trợ tương ứng của nó). Khoảng cách giữa 2 siêu phẳng hỗ trợ được gọi là lề. Siêu phẳng tối ưu (nằm giữa 2 siêu phẳng hỗ trợ) tìm được từ 2 tiêu chí là cực đại hóa lề (lề càng lớn, mô hình phân lớp càng an toàn) và cực tiểu hóa lỗi. Vấn đề dẫn đến việc giải bài toán quy hoạch toàn phương (1):

$$\min \Psi(w, b, z) = (1/2)\,||w||^2 + c\sum_{i=1}^{m} z_i$$

$$s.t. \quad (1)$$

$$y_i(w.x_i - b) + z_i \geq 1$$

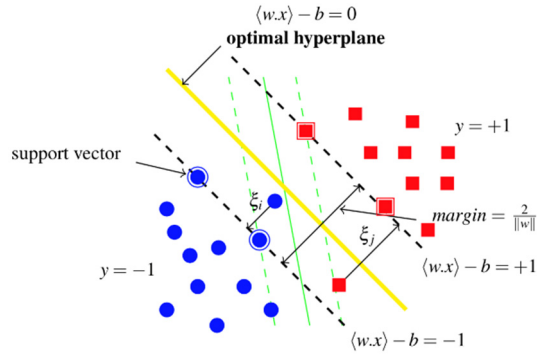
$$z_i \geq 0 \quad (i=1,m)$$

hằng $c > 0$ sử dụng để chỉnh độ rộng lề và lỗi

Giải bài toán quy hoạch toàn phương (1), thu được (w, b). Phân lớp phần tử x dựa vào biểu thức $sign(w.x - b)$.

Mặc dù giải thuật SVM cơ bản chỉ giải quyết được bài toán phân lớp tuyến tính, tuy nhiên nếu ta kết hợp SVM với phương pháp hàm nhân, sẽ cho phép giải quyết lớp các bài toán phân lớp phi tuyến (Cristianini & Shawe-Taylor, 2000).

(Platt, 1998) chỉ ra rằng các giải thuật huấn luyện được đề xuất trong (Boser *et al*., 1992), (Chang & Lin, 2011), (Osuna *et al*., 1997), (Platt, 1998) có độ phức tạp tính toán lời giải bài toán quy hoạch toàn phương (1) tối thiểu là $O(m^2)$ trong đó m là số lượng phần tử được dùng để huấn luyện. Điều này làm cho giải thuật SVM không phù hợp với dữ liệu lớn.



Hình 1: Phân lớp tuyến tính với máy học SVM