ENHANCEMENT OF NEW SMOOTH SUPPORT VECTOR MACHINES FOR CLASSIFICATION PROBLEMS

SANTI WULAN PURNAMI

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Doctor of Philosophy in Computer Science

Faculty of Computer System and Software Engineering
UNIVERSITI MALAYSA PAHANG

JUNE 2011

ABSTRACT

Research on Smooth Support Vector Machine (SSVM) for classification problem is an active field in data mining. SSVM is reformulation of standard Support Vector Machines (SVM). In SSVM, smoothing technique must be applied to convert constraint optimization to the unconstraint optimization problem since the objective function of this unconstraint optimization is not twice differentiable. The smooth function is used to replace the plus function to obtain a smooth support vector machine (SSVM). To get more accuracy performance, Multiple Knot Spline SSVM (MKS-SSVM) is proposed. MKS-SSVM is a new SSVM which used multiple knot spline function to approximate the plus function instead the integral sigmoid function in SSVM. To obtain optimal accuracy results, Uniform Design method is used to select parameter. The performance of the method is evaluated using 10-fold cross validation accuracy, confusion matrix, sensitivity and specificity. To evaluate the effectiveness of our method, an experiment is carried out on four medical dataset, i.e. Pima Indian diabetes dataset, heart disease, breast cancer prognosis, and breast cancer diagnosis. The results of this study showed that MKS-SSVM was effective to diagnose medical dataset and this is promising results compared to the previously reported results. SSVM algorithms are developed for binary classification. However, in many real problems data points are discriminated into multiple categories. Hence, MKS-SSVM is extended for multiclass classification. Two popular multiclass classification methods One against All (OAA) and One against One (OAO)) were used to extend MKS-SSVM. Numerical experiments show that the classification accuracy of OAA and OAO method are competitive with each other and there is no clear superiority of one method over another. While the computation time, the OAO method is lower than the OAA method on three dataset. This indicated that the OAO method is usually more efficient than the OAA. In the final part, the reduced support vector machine (RSVM) was proposed to solve computational difficulties of SSVM in large dataset. To generate representative reduce set for RSVM, clustering reduced support vector machine (CRSVM) had been proposed. However, CRSVM is restricted to solve classification problems for large dataset with numeric attributes. In this research, an alternative algorithm, *k*-mode RSVM (KMo-RSVM) that combines RSVM and *k*-mode clustering technique to handle classification problems on categorical large dataset and *k*-prototype RSVM (KPro-RSVM) which combine *k*-prototype and RSVM to classify large dataset with mixed attributes were proposed. In our experiments, the effectiveness of KMo-RSVM is tested on four public available dataset. It turns out that KMo-RSVM can improve speed of running time significantly than SSVM and still obtained a high accuracy. Comparison with RSVM indicates that KMo-RSVM is faster, gets smaller reduced set and comparable testing accuracy than RSVM. From experiments on three public dataset also show that KPro-RSVM can tremendously reduces the computational time and can handling classification for large mixed dataset, when the SSVM method ran out of memory (in case: census dataset). The comparison with RSVM indicate that the computational time of KPro-

RSVM less than RSVM method, and obtained testing accuracy of KPro-RSVM a little decrease than RSVM.

ABSTRAK

Penyelidikan Mesin Vector Sokongan Licin ((SSVM) adalah bidang yang aktif dalam pelombongan data. SSVM adalah perumusan semula dari Mesin Vektor Sokongan (SVM). Dalam SSVM, teknik pelicinan diterapkan untuk menukarkan pengoptimuman berkekangan dengan masalah pengoptimuman tidak berkekangan karena fungsi tujuan dari pengoptimuman tidak berkekangan tidak dibezakan. Fungsi *pelicinan* digunakan untuk menggantikan fungsi *plus (plus function)* sehingga disebut Mesin Vector Sokongan Licin (SSVM). Untuk mendapatkan ketepatan yang lebih baik, *Multiple Knot Spline* SSVM (MKS-SSVM) dicadangkan untuk masalah pengkelasan. MKS-SSVM adalah SSVM baru yang menggunakan *Multiple Knot Spline* untuk menganggarkan fungsi *plus* menggantikan fungsi integral sigmoid dalam SSVM. Untuk mendapatkan hasil ketepatan yang optimum, kaedah Uniform Design digunakan untuk memilih parameter. Prestasi MKS-SSVM dinilai menggunakan *10-fold cross validation*, *confusion matrix*, *sensitivity* dan *specitivity*. Untuk menilai keberkesanan kaedah ini, percubaan dilakukan pada empat dataset perubatan, iaitu dataset diabetes, penyakit jantung, prognosis kanser payudara, dan diagnosis kanser payudara. Keputusan kajian ini menunjukkan bahawa MKS-SSVM berkesan untuk mendiagnosis dataset perubatan dan ini sangat menjanjikan hasil berbanding dengan keputusan yang dilaporkan sebelum ini. Algoritma SSVM dibangunkan untuk pengkelasan perduaan. Walau bagaimanapun, dalam banyak masalah sebenar data didiskriminasi ke dalam berbilang kategori. Oleh itu, MKS-SSVM dilanjutkan untuk pengkelasan berbilang kategori. Dua kaedah pengkelasan yang popular iaitu *One Agains All* (OAA) dan *One Against One* (OAO) digunakan untuk membangun MKS-SSVM. Dari eksperimen menunjukkan bahawa kaedah ketepatan klasifikasi OAA dan OAO bersaing antara satu sama lain dan tidak ada keunggulan yang jelas dari satu kaedah di atas yang lain. Dalam bahagian akhir, Reduced Support Vector Machines (RSVM) telah dicadangkan untuk menyelesaikan masalah pengiraan SSVM dalam dataset yang besar. Untuk menjana *reduce set* untuk RSVM, clustering reduced support vector machine (CRSVM) telah dicadangkan. Walau bagaimanapun, CRSVM adalah terhad untuk menyelesaikan masalah pengelasan untuk dataset besar dengan sifat-sifat angka. Dalam kajian ini, algoritma alternatif, *k*-mode RSVM (KMo-RSVM) yang menggabungkan RSVM dan *k-modes* clustering teknik untuk menangani masalah pengelasan pada dataset kategori yang besar dan *k*-prototaip RSVM (KPro-RSVM) yang menggabungkan *k*-prototaip dan RSVM untuk mengelaskan dataset besar dengan sifat-sifat campuran telah dicadangkan. Dalam percubaan kami, keberkesanan KMo-RSVM diuji pada empat dataset. Ternyata KMo-RSVM dapat meningkatkan kelajuan masa secara signifikan dari SSVM dan masih memperoleh ketepatan yang tinggi. Perbandingan dengan RSVM menunjukkan bahawa KMo-RSVM lebih cepat, mendapatkan set yang lebih kecil dan mengurangkan ketepatan ujian setanding dari RSVM. Dari percubaan pada tiga dataset awam juga menunjukkan bahawa KPro-RSVM dapat mengurangkan masa pengkomputeran secara signifikan dan dapat menangani pengkelasan untuk dataset campuran, ketika kaedah SSVM kehabisan memori (dalam hal: dataset *census.*