ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Τζιράκης Παναγιώτης

Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτ. Εργασίας: Αναπλ. Καθηγητής, Ι. Τσαμαρδίνος

Πέμπτη, 15/10/2015, 15:00

Αίθουσα Κ206, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

" Υλοποίηση Αλγορίθμων Επιλογής Μεταβλητών για Μεγάλο Όγκο Δεδομένων "

ΠΕΡΙΛΗΨΗ

Στις μέρες μας, υπάρχει εκθετική αύξηση των δεδομένων τόσο στον αριθμό των δειγμάτων όσο και στον αριθμό των μεταβλητών, με τον μέγεθος τους να φτάνει την κλίμακα των terabyte. Αυτός ο όγκος δεδομένων μπορεί να βρεθεί σε πολλές εφαρμογές της μηχανικής μάθησης όπως στην ανάκτηση πληροφοριών, κατηγοριοποίηση κειμένου και ανάκτηση εικόνων. Παρόλο που τέτοιου είδους δεδομένα είναι συχνά σήμερα, κλασσικοί αλγόριθμοι μηχανικής μάθησης δεν μπορούν να τα διαχειριστούν.

Μια πολύ σημαντική μέθοδος στη μηχανική μάθηση είναι η επιλογή μεταβλητών που προσπαθεί να επιλέξει τις μεταβλητές που είναι πιο προβλεπτικές σε ένα σετ δεδομένων. Η επιλογή μεταβλητών είναι σημαντική καθώς μειώνει τις διαστάσεις των δεδομένων, αφαιρεί άσχετες μεταβλητές, αυξάνει την επίδοση ενός ταξινομητή και βοηθάει στην

καλύτερη κατανόηση των δεδομένων. Με την αύξηση του όγκου των δεδομένων η απόδοση των κλασσικών αλγορίθμων επιλογής μεταβλητών μειώνεται αισθητά.

Για να λυθούν προβλήματα απόδοσης, το μοντέλο Map-Reduce έχει προταθεί. Τα δεδομένα πλέον μπορούν να επεξεργαστούν παράλληλα σε ένα σύμπλεγμα υπολογιστών και οι αλγόριθμοι μηχανικής μάθησης μπορούν να τροποποιηθούν έτσι ώστε να είναι σε θέση να επεξεργαστούν μεγάλο όγκο δεδομένων.

Σε αυτή την εργασία ασχοληθήκαμε με την υλοποίηση ενός αλγορίθμου επιλογής μεταβλητών για μεγάλο όγκο δεδομένων. Πιο συγκεκριμένα, χρησιμοποιήσαμε το μοντέλο Map-Reduce για να παραλληλοποιήσουμε τον αλγόριθμο Max Min Parent and Children (MMPC) έτσι ώστε να μπορεί να διαχειριστεί μεγάλο όγκο δεδομένων. Ο αλγόριθμος προσπαθεί ευριστικά, με τη χρήση τεστ ανεξαρτησίας, να βρει εξαρτήσεις μεταξύ μεταβλητών. Σε αυτή την εργασία δείχνουμε πως παραλληλοποιήσουμε δύο τεστ ανεξαρτησίας, που μπορούν να διαχειριστούν κατηγορικές και συνεχείς μεταβλητές, χρησιμοποιώντας το μοντέλο Map-Reduce. Τέλος, χρησιμοποιήσαμε μια μέθοδο με την οποία ο MMPC μπορεί να χρησιμοποιηθεί με οποιοδήποτε τεστ ανεξαρτησίας.

Για να αξιολογήσουμε τον αλγόριθμο χρησιμοποιήσαμε δεδομένα που περιέχουν διαφορετικό αριθμό δειγμάτων και μεταβλητών. Η αξιολόγηση έδειξε ότι ο αλγόριθμος μας κλιμακώνεται καλά όταν αλλάζει ο αριθμός των δειγμάτων και ο αριθμός των κόμβων στο σύμπλεγμα υπολογιστών. Τέλος, η απόδοση του αλγορίθμου είναι συγκρίσιμη με την απόδοση άλλων αλγορίθμων επιλογής μεταβλητών.

Tsirakis Panagiotis

M.Sc. Thesis

Computer Science Department

University of Crete

Master's Thesis Supervisor: Associate Professor I. Tsamardinos

Thursday, 15/10/2015, 15:00

Room K206, Computer Science dept., University of Crete

“Implementing Feature Selection Algorithms for Big Data”

ABSTRACT

In recent years, data has an exponential growth in both the number of instances and the number of features, which brings their scale to the level of terabytes. These amounts of data can be found in many machine learning applications like information retrieval, text categorization and image retrieval. Although such amounts of data are very frequent nowadays, classical machine learning algorithms cannot handle them.

A very important task in machine learning is feature selection and its task is to select the most informative features in a dataset. Feature selection is effective in reducing dimensionality, removing irrelevant features, increasing performance of a learner, and improving our understanding of the data. With the increase of the volume of the data the usability of classical feature selection algorithms significantly deteriorates.

To solve scalability problems, the Map-Reduce model has been proposed. With this model the data can be processed in parallel in a cluster, and so machine learning algorithms can now be altered in order to process terabytes of data.

In this thesis we were concerned with the implementation of a feature selection algorithm for big data. More particularly, we used the Map-Reduce model to parallelize the Max-Min Parent and Children (MMPC) algorithm in order to be able to handle big data. MMPC tries, heuristically, with the use of independent tests, to find dependencies among the features. For this thesis we show how two independence tests that can handle categorical and continuous features, can be used with the Map-Reduce model. Finally, we also use a method so that MMPC can be used with any independence test using the Map-Reduce model.

To evaluate our algorithm, we experimented with datasets that contained different number of instances and features. The experimental evaluation showed that our algorithm scales well with these datasets when varying the number of instances and the number of nodes in the cluster. Moreover, the performance of the algorithm is comparable to other feature selection algorithms.