Fachhochschule Brandenburg
Fachbereich Informatik und Medien

Integration von Data-Mining-Methoden zur Analyse der Daten aus dem astrophysikalischen Experiment LOPES

Masterarbeit, vorgelegt von Marcin Franc

Aufgabenstellung:

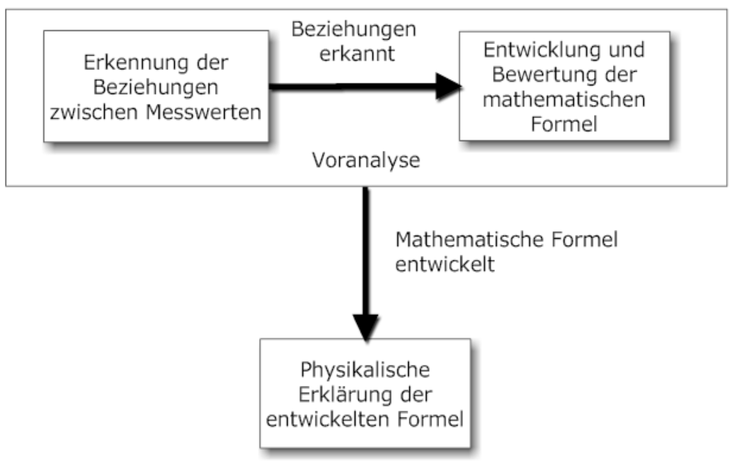
Das Ziel der Arbeit ist die Evaluierung der Data-Mining-Methoden, die zur Analyse der LOPES-Daten benutzt werden könnten, die erfolgversprechendsten zu wählen und diese in eine zu entwickelnde Software zu integrieren. Darüber hinaus muss die Software die von den Forschern genannten Anforderungen erfüllen.

Vorstellung des LOPES-Experiments:

Das Experiment findet an dem Karlsruher Institut für Technologie (KIT) statt und versucht im Allgemeinen die Beschaffenheit der kosmischen Strahlung zu beschreiben. Diese Strahlung besteht aus Teilchen mit einer Energie von mehr als 10^{10} eV, die mit der Erdatmosphäre kollidieren und dabei einen sog. Luftschauer aus Sekundärteilchen auslösen. Diese Sekundärteilchen wiederum werden im Erdmagnetfeld durch die Lorentzkraft abgelenkt und emittieren dabei einen Radiopuls, der mit Experimenten wie LOPES gemessen wird. Die Radio-Emission von diesen Teilchen wurde zuerst von Jelley im Jahr 1965 entdeckt.

Status quo:

Die bisherigen Aufgaben der Physiker kann man in zwei Gruppen unterteilen. Die erste Aufgabengruppe besteht aus der Erkennung potenzieller Beziehungen zwischen den Messwerten sowie der Entwicklung und Bewertung einer mathematischen Formel, mit der diese Beziehungen beschrieben werden können. Wie bereits erwähnt, ist diese Situation einfacher, da man weiß, was gesucht werden soll. Die Aufgaben aus der zweiten Gruppe umfassen die genaue Erklärung der erhaltenen Ergebnisse mit der existierenden physikalischen Theorie.



Schema der bisherigen Analyse-Aktivitäten

Der erste Teil der Analyse besteht aus der Entwicklung kleinerer Stücke Software, die bestimmte mathematische und statistische Operationen durchführen sollen und als Ausgabe die Formel liefern, die später aus physikalischer Sicht analysiert werden kann.

Verbesserungsmöglichkeiten:

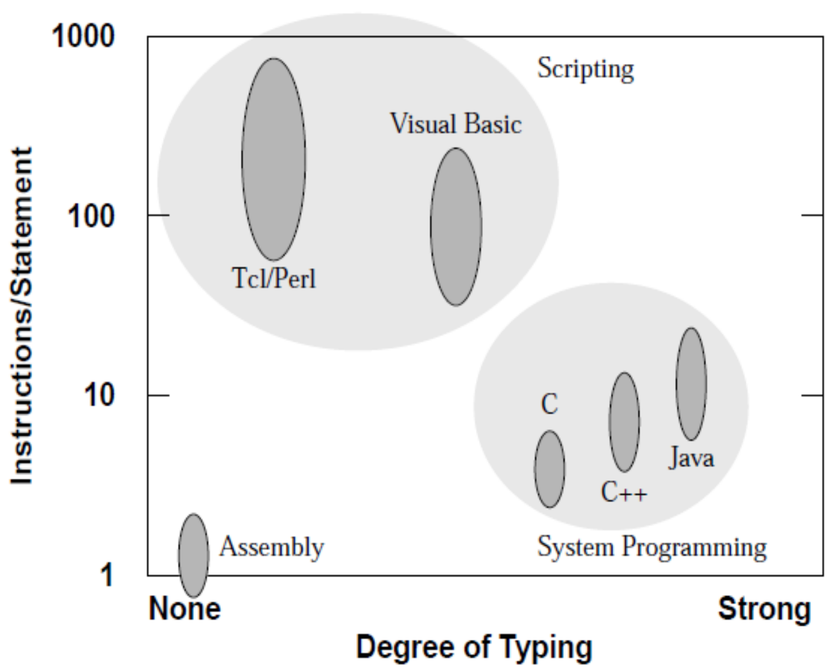
Es existieren ein paar Probleme mit der bisherigen Lösung. Unter anderem, dass ein Großteil der Quellcode-Teile meistens „Aufgaben-orientiert“ geschrieben wurde, sodass sie für spätere ähnliche Aufgaben nutzlos sind. Diese Schwierigkeiten induzieren, dass man relativ viel Zeit und Ressourcen braucht, um eine Lösung zu finden. Die generell-formulierte Verbesserungsidee ist, sowohl die Methoden des Data Mining zu benutzen, um die Suche nach Abhängigkeiten zwischen den Messwerten teilweise zu automatisieren, als auch eine hochqualitative Software zu entwickeln, die später auch eventuell ausgebaut werden könnte.

Die entwickelte Software:

Eine der wichtigsten Anforderungen an die Software ist, dass sie einfach konfiguriert und ausgebaut werden können soll. Um das zu erreichen, wurde eine Bibliothek von Python-Klassen aufgebaut, die sowohl für die Entwicklung der Hauptapplikation benutzt wurde als auch das einfache Schreiben eigener Skripte erlaubt. Da die Hauptapplikation in Python geschrieben wurde, könnte ihr Quellcode selbst von den Forschern modifiziert werden, um die gewünschte Funktionalität zu erreichen. Man darf nicht vergessen, dass die Zielgruppe der Software sowohl aus Entwicklern als auch technisch fortgeschrittenen Benutzern besteht, die eine genaue Vorstellung davon haben, was sie von der Software erwarten. Die Möglichkeit der beliebigen Modifikationen mit einer Skriptsprache kann diese Erwartungen erfüllen. Die entwickelte Applikation ist keine kompilierte „Black Box“-Software, sondern liegt im Quellcode vor, der später modifiziert werden kann.

Vorteile der Skriptsprachen:

Die wichtigsten Vorteile der Skriptsprachen, die eine schnelle und einfache Entwicklung der Programme (Skripte) erlauben, sind dynamische Typisierung und die große Anzahl der Maschinenbefehle pro Zeile.



Vergleich verschiedener Programmiersprachen auf Grund der Anzahl der Maschinenbefehle pro Anweisung und Grad der Typisierung,

Auswahl der Data-Mining-Methoden:

Die Literatur schlägt immer vor, dass die einfachsten Lösungen sehr oft ausreichende Resultate liefern. Deswegen wird die Auswahl der Methoden, die in der entwickelten Software benutzt werden, unter Berücksichtigung des KISS-Prinzips durchgeführt.

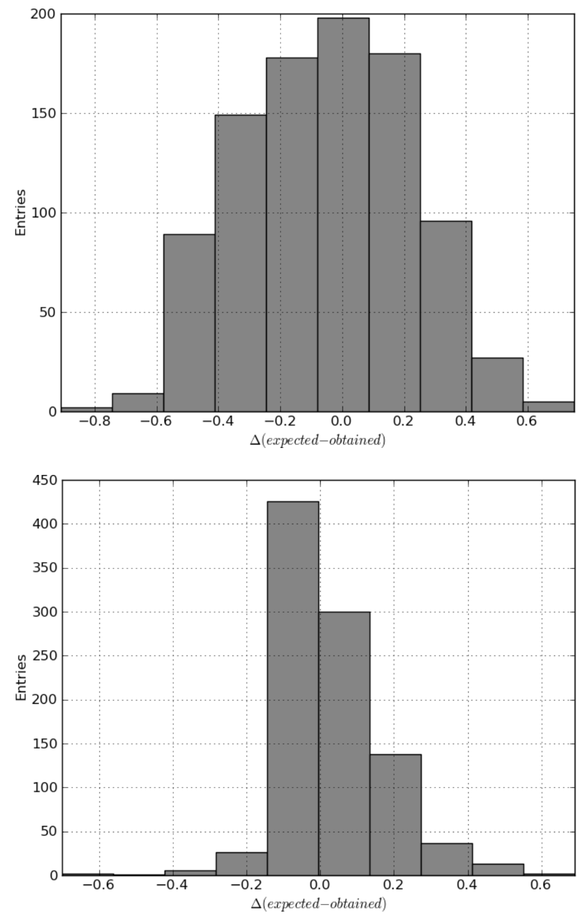
Die Methoden, die für die Zwecke der Arbeit gewählt werden, waren unter anderem:

- SVM
- genetische Programmierung

Man kann diskutieren, ob genetische Programmierung selbst eine Art von Data Mining ist. Verschiedene Bücher klassifizieren evolutionäre Algorithmen als eine der Methoden des Data Mining, treffen aber über genetische Programmierung keine Aussage. Das Ziel der Methode ist aber kongruent mit der Definition des Data Mining und deswegen wird für den Zweck dieser Arbeit als Voraussetzung angesehen, dass die Methode wirklich als eine der Data-Mining-Methoden betrachtet werden kann.

Ergebnisse:

Die Software wurde zu diesem Zeitpunkt u. a. benutzt, um die Performance von Data-Mining-Lösung und bisher im Experiment verwendeter Formel miteinander zu vergleichen. Die dargestellten Ergebnisse basieren auf der Formel zur Berechnung der Radiopulshöhe auf Grund der geschätzten Primärenergie. Diese Formel wurde umgestellt, um die Primärenergie zu berechnen. Das Ergebnis wurde mit dem Referenzwert aus dem KASCADE-Experiment verglichen. Mit den selben Daten wurde eine Kreuzvalidierung für die Methode der genetischen Programmierung durchgeführt. Die Ergebnisse dieser zwei Verfahren werden mit der folgenden Abbildungen präsentiert (unteres Bild).



Betreuer: Prof. Dr. Jochen Heinsohn
Prof. Dr. Johannes Blümer

Abgabe: 18.08.2010