**COMP500 RESEARCH PROPOSAL**

**Title: Exploring the application of the Particle Swarm Optimization algorithm in the area of streaming data analytics.**

Matthew XXXXX

October-21-18

Student No. XXXXX

**Abstract:** I propose to extend a Particle Swarm Optimization algorithm to improve its effectiveness at solving classification problems with streaming data in dynamic environments.

**Background:** Data mining or Knowledge discovery in databases (KDD) is an automated process of extracting useful, comprehendible, interesting and previously unknown patterns from large amount of data using techniques such as machine learning, statistical computations and computational intelligence [1]. As the volume, velocity, variety and veracity of data collection continually increase there is an ongoing need to improve the machine learning techniques to make use of the data.

Evolutionary computing (EC) encompasses a class of optimization and classification algorithms that rely on the Darwinian principals such as biological evolution and natural selection. Particle Swarm Optimization (PSO) is an algorithm that mimics the behavior of schooling species such as fish or flocking birds. In a PSO algorithm, candidate solutions, each of which is termed a particle, are in the population pool and are moving through the problem hyperspace. In each iteration, every particle is evaluated by a defined fitness function. Based on its fitness value, each particle is then stochastically accelerated towards its previous best known position and also the best known position of the swarm [2].

**Motivation:** Many of the most effective machine learning algorithms such as deep neural networks achieve their efficiency through extensively training complex algorithmic architectures with an abundance of data. An important issue of ongoing research is transfer learning, which focuses on being able to extract useful information from existing models and applying them to new problems, thereby reducing the cost of retraining a new model from scratch. As PSO has an intrinsic ability to handle dynamic environments, PSO algorithms can be designed to reduce the amount of retraining necessary for different problem environments. In my research project I would seek to enhance the capability of a PSO algorithm to perform classification tasks in dynamic environments with streaming data. As an illustrative example, consider the problem of predicting high frequency stock market shocks. PSO can be used to evolve solutions of the sign and amplitude of the next transaction in the stock market, and as each new actual transaction is realized, a new solution particle based on the latest shock, with a unique influential weight can be added to the solution pool, so that a new set of solutions can be trained with reduced computation in the new data environment.

While there have been several theoretical papers demonstrating the effectiveness of various modifications of a PSO algorithm, a few of them have been applied in practice. Current limitations of the algorithm are premature convergence and that the performance is dependent on choosing the right parameters for the algorithm, which need to be set manually [3].

**Objective:** I plan to research the most effective versions of PSO algorithms and apply them to a variety of problems that have both a dynamic environment and streaming data, from predicting stock market shocks to predicting the next location ‘check-in’ of a user in a social network. I will also research solutions to such problems given by other ML algorithms to evaluate the general effectiveness of PSO in these scenarios. After evaluating the shortcomings of current PSO algorithms I hope to implement changes such as continually steaming new solutions into the candidate pool, and using multiple swarms with ideas from transfer learning to reduce learning time and improve accuracy. Finally, the implementation will be retested and the results can be compared and analyzed.

**Project Scope (data and tasks):** I will be using public data sets for carrying out my tasks on clustering text data using PSO.

**Deliverables:** Poster, Thesis Defense (not needed for 499), Final Report

**Timeline:**

|  |  |
| --- | --- |
| July-Oct | Research PSO implantations, limitations. Retrieve datasets to test algorithms on. Research existing solutions to problems presented in datasets. |
| Oct | Write background literature review and update proposal. |
| Nov 1 | Literature review due with proposal. |
| Nov-Dec | Implement existing PSO algorithms and test with datasets. Test other ML algorithms with data sets. |
| Jan-Feb | Analyzing the results of existing algorithms, iterate through modifications of PSO and record performance effects. |
| Feb-Mar | Write thesis document. Prepare presentation. |

Presentation: End of last week of winter classes at the latest.

Thesis: Due two weeks before the presentation.

Other work: Includes a poster for the Creating Computing event and the software code implemented as a part of the project. This also includes the data collected for evaluation purposes. Due Friday the last week of class at the latest.

**References**

Edelstein, H. *Introduction to data mining and knowledge discovery*, 3rd edn. Two Crows Corp. 1999

Kennedy, James F., Russell C. Eberhart, and Yuhui Shi. *Swarm Intelligence*. San Francisco: Morgan Kaufmann, 2001

S. Aote. *A Brief Review on Particle Swarm Optimization: Limitations & Future Directions*, IJCSE, Vol 2 no.5 2013.