**Assignment (CS-560)**

**By Vidhi Shah**

**Student ID: 16178723**

**Critique on Paper-1**

**Data Science and Prediction**

**Summary:**

Research on manipulating big data in a large business in terms of data science has drawn attraction of many people nowadays. Combining basic knowledge of Bayesian statistics, how data is stored and manipulated by computers and how to find correlation and causation among data are the requirements to work with. The key points of this paper explain the importance of predictive accuracy in machine learning and knowledge discovery in database.

Machine learning skill is the most important requirement for researchers. Based on this skill one can get automatic resolution which is depending on accurate prediction. The word machine learning is connected with two more words: depiction and generalization. All machines learning system has one of the parts as depiction of data instance. On other side unseen data instance comes under generalization.

Big data network topology consist of four units: user interaction and learning system, network traffic recording system, HDFS and cloud computing storage system. Using this topology, machine learning (ML) techniques find best suitable algorithm and thus facilitating desire output.

**Critique:**

Unlike relational data big data is having structure and unstructured data as well. So big data has a complex data structure and it is facing number of issues related to data and manipulation. Big data consist of variety of class, different nonrelative’s dataset and different distributed data sources which creates complexity and results in lack of accuracy and takes longer time to proceed data.

From the labeled database, data is fetched using certain criteria in machine learning technique. But in these ML techniques there are few drawbacks. Firstly, it is not suitable for today’s multiple learning tasks. Moreover, fast growing dynamic dataset results in false results. On top of this, Classification may not be strong over distributed database. For classification of network data, classification algorithms are used. In all machine learning (ML) techniques associated algorithms, nowadays Support Vector Machine (SVM) is highly demanded for classification of network traffic but it is not cost effective. Though it gives accurate classification, it is a big issue to make an alternative of SVM or to modify this in accordance to make it economical.

**Questions:**

1. The paper talks about machine learning technique. But in that highly accurate algorithm like SVM is very costly. How to minimize the cost and improve SVM?
2. The paper mentioned that it is inexpensive data storage though we required large space to store big data of number of different field. So how to satisfy additional storage place for data from public cloud for processing big data?
3. Big data is largely used for business where we need to have a confidential data. Regarding to that how can we implement robust security system on big data?