**COMP\_6600 Artificial Intelligence**

**Course Project Proposal**

**Solve E-commerce Order Batch problem by Clustering**

Chengfei Wang

**1. Introduction**

E-commerce booming brings challenge to traditional management of logistic. How to efficiently satisfy the requirements from customer becomes the priority of the company like amazon. Improve efficient of process of orders by grouping of similar ones into one batch is a classic method used in company. But it is impossible to manually do the task for large amount of orders (>10,000) meanwhile take all the factors of warehouse inventory status, locations of distribution centers and costumers, capacity and time constrains into account. A machine learning clustering solution should be proposed to solve the automation problem of order batch optimization problem, therefore the best or the second best efficient can be achieved.

**2. Problem Description**

The logistics systems usually consist of warehouses, distribution centers and transportation vehicles. The stock keeping units (SKU) are the inventory in the warehouses catalog, and the catalogs of different warehouse may overlaps with each other. The distribution center collects items from warehouses according to the given set of orders, and delivers them to customers in local region by transportation vehicles. For the seller, it draws great effort to lower the cost of operation in the sorting process, therefore the grouping of orders is the first method we can think of. The cluster of orders must minimize the sorting cost at warehouses and travel cost of vehicles while satisfying a certain constrains, both time and capacity.

Formally, we have a sets of the warehouses ***W*** and the distribution centers ***C,*** where ***W***∪***C*** = ***V*** and ***W***∩***C***= ∅, the number of distribution centers determines the number of clusters, denoted by *k* = |***C***|. The distance between vertex *vi* and *vj* is expressed by *dij* > 0, *vi*, *vj*∈***V***. ***V*** and the set of distance ***E*** form a undirected graph *G* = {***V***, ***E***}

The input set ***X*** = {**x**1, **x**2, …, **x***N*} is the set of on-line orders received by the retailer and each element **x***i*∈R*d*represents the attributes of the *i*-th order.are denoted by and ***C***, Each order is identified by multiple attributes as follows.

* Invoice ID: the unique identification of an order
* Commodities: the collection of items which are recorded by their unique stock codes and quantities
* Customer ID: the identification of customer who places the order
* Deadline: the time before which the order must arrive at the customer

Define the distance function *u*: R*d*×R*d* → R+ be a function between any two points in R*d*. Optimization goal is to find a disjoint *k*-partitioning of ***X*** and a corresponding set of *k* cluster representatives *M* = in R*d* for a given *k* such that the clustering objective function

is minimized under the load balance constraint (, , for a given *m* with *mk* *N*.)

Given above conditions, this project focused on solving three challenges:

(1) How to precisely define a meaningful distance between every two orders.

(2) How to deal with the balancing constraints due to time and capacity constrains.

(3) How to evaluate the result of the proposed clustering method.

**3. Dataset**

An on-line retail dataset in UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/datasets/online+retail>) will be used to train our machine learning model. The dataset data set contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered on-line retail. The company mainly sells unique all-occasion gifts. The customers of the company are wholesalers.