ERCIM "Alain Bensoussan" Fellowship Scientific Report

Fellow:	**Xiangliang Zhang**
Visited Location:	Department of Computer and Information Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
Duration of Visit:	April 1st, 2010 ~ August 31st, 2010

I - Scientific activity

(1 page at maximum)

During my 5-month fellowship period, I theoretically studied the issues of generating a specified number of clusters in the clustering algorithm called Affinity Propagation (AP). Motivated by the automatic text abstraction, I applied the proposed clustering method for extracting a given number of keywords from a text document.

AP is a clustering algorithm that provides optimal guarantee about minimizing the clustering distortion, compared to *k*-medoids. In counterpart for this guarantee, AP is limited by its quadratic computational complexity, and by the fact that it does not allow directly specifying the number of clusters. Instead, the number of clusters produced by AP is implicitly controlled by a user-defined parameter, which is the self-confidence for each item to be an *exemplar*. However, in many application domains, prior knowledge indicates the number of clusters in the data. Clustering algorithms are thus required to generate a desired number of clusters. For example, *k*- medoids has an advantage in directly generating a specified number of clusters.

In our study, we aim at modifying AP to directly provide a given number of clusters while remaining all its advantages in clustering, e.g., the minimum distortion. We call this proposed method *k*-AP. Through adding a constraint function to limit the number of clusters to be a requested one, the confidence of one data item to be an exemplar in *k*-AP is automatically self-adapted, while the confidence is a parameter specified by users in AP.

We validate *k*-AP on several benchmark data sets from UCI Machine Learning Repository. The experimental results show that *k*-AP improves on the state of the art w.r.t. the distortion minimization and higher clustering purity and less computational complexity.

k-AP is also applied for keywords extraction from Reuters-21578 data set, which includes 21,578 documents assembled and indexed with 135 categories. The experimental results show that the extracted keywords correspond well to the content of the documents.

II- Publication(s) during your fellowship

Please insert the title(s), author(s) and abstract(s) of the published paper(s). You may also mention the paper(s) which were prepared during your fellowship period and are under reviewing.

***K*-AP: Generating Specified K Clusters by Efficient Affinity Propagation**. Xiangliang Zhang, Wei Wang, Kjetil Norvag, and Michele Sebag. **Accepted** by IEEE International Conference on Data Mining (ICDM) 2010.

Abstract: The Affinity Propagation (AP) clustering algorithm proposed by Frey and Dueck (2007) provides an understandable, nearly optimal summary of a data set, through message passing among all pairs of data items. However, it suffers two major shortcomings: i) the number of clusters is vague with the user-defined parameter called self-confidence, and ii) the quadratic computational complexity. When aiming at a given number of clusters due to prior knowledge (e.g., number of classes in a supervised context), AP has to be launched many times until an appropriate setting of self-confidence parameter is found by a bisection method. The re-launched AP increases the computational cost by 1 order of magnitude. In this paper, we propose an algorithm, called K-AP, to exploit the immediate results of K clusters by introducing a constraint in the process of message passing. Through theoretical analysis and experimental validation, K-AP was shown to be able to directly generate K clusters as user-defined, with a negligible increase of computational cost compared to AP. In the meanwhile, K-AP preserves the clustering quality as AP in terms of the distortion (sum of the squared distance between each data item and its assigned cluster center). It is more effective than k-mediods w.r.t. the distortion minimization and higher clustering purity.

Adaptively detecting changes in autonomic grid computing. Xiangliang Zhang, Cecile Germain and Michele Sebag. **In: Proceedings** of 11th ACM/IEEE International Conference on Grid Computing (Grid 2010), workshop on Autonomic Computational Science.

Abstract: Detecting the changes is the common issue in many application fields due to the non-stationary distribution of the applicative data, e.g., sensor network signals, web logs and grid running logs. Toward Autonomic Grid Computing, adaptively detecting the changes in a grid system can help to alarm the anomalies, clean the noises, and report the new patterns. In this paper, we proposed an approach of self-adaptive change detection based on the Page-Hinkley statistic test. It handles the non-stationary distribution without the assumption of data distribution and the empirical setting of parameters. We validate the approach on the EGEE streaming jobs, and report its better performance on achieving higher accuracy comparing to the other change detection methods. Meanwhile this change detection process could help to discover the device fault which was not claimed in the system logs.

Self-adaptive Change Detection in Streaming Data with Non-stationary Distribution. Xiangliang Zhang and Wei Wang. **Accepted** by International Conference on Advanced Data Mining and Applications (ADMA2010).

Abstract: Non-stationary distribution, in which the data distribution evolves over time, is a common issue in many application fields, e.g., intrusion detection and grid computing. Detecting the changes in massive streaming data with non-stationary distribution helps to alarm the anomalies, clean the noises, and report the new patterns. In this paper, we employ a novel approach for detecting changes in streaming data with the purpose of improving the quality of modeling the data streams. The self-adaptability of the novel approach enhances the effectiveness of modeling data streams by timely catching the changes of distributions. We

validated the approach on an online clustering framework with a benchmark KDDcup 1999 intrusion detection data set as well as a real-world grid data set. The validation results demonstrate its better performance on achieving higher accuracy and lower percentage of outliers comparing to the other change detection approaches.

III -Attended Seminars, Workshops, and Conferences

Please identify the name(s), date(s) and place(s) of the events in which you participated during your fellowship period.

NONE

IV – Research Exchange Programme (12 month scheme)

Please identify the name(s), date(s) and place(s) of your Research Exchanges during your fellowship period and detail them.

July 1st to July 31th, 2010
Visited Interdisciplinary Centre for Security, Reliability and Trust (SnT Centre), University of Luxembourg, Luxembourg (hosted by Prof. Bjorn Ottersten).