Master in Artificial Intelligence (UPC-URV-UB)

Master's Thesis Proposal

General Information

Master's Thesis Title:	Automatic technique using data fusion dealing with the data absence
Publication Date:	March 2011
Expiry Date:	June 2011
Orientation:	☐ professional ☐ research
M.Sc. Th. Advisor:	Tomàs Aluja-Banet
M.Sc. Th. Advisor's Dept. & University:	the Department of Statistics and Operations Research, UPC
M.Sc. Th. Advisor e-mail:	tomas.aluja@upc.edu
Observations:	
Student's Name: (if already known)	Peng Si

M.Sc. Thesis Description

Main issues / Brief Description:

In this thesis paper the student will aim at developing a software tool which is R project for automatically dealing with the Data Fusion problems. Data fusion, is defined as the use of techniques that combine data from multiple sources and gather that information in order to achieve inferences, which will be more efficient and potentially more accurate than if they were achieved by means of a single source.

Detailed Description:

As its name implies, the project aims at developing a tool software for automatic data fusion. In this tool we called Graft. Graft applies the knn methodology and implement local computation. The design will be done in two completely separate modules,which enable their use independently. The first module will search individuals(donors) who have more in common with individuals(recipient) who has missing information and will be estimated. So Therefore, donor and recipient are placed in a common space, and each recipient will look for other individuals k similar among all donors (k Nearest Neighbours). The second module will have to impute the missing information from the neighborhood relations, between donors and recipients which have been found previously.

To achieve this goal we must achieve other objectives listed below:
1. Implement a fast search algorithm of the k nearest neighbors. For achieve this goal we will analyze the different fast knn search algorithms.
2. Modify the search algorithm selected to be the knn to apply restrictions on the neighbors.
3. Designing and implementing a reciprocal neighbors search algorithm . Search algorithms are reciprocal neighbors only if you have the k nearest neighbors, so we will find a new algorithm which may take advantage of this information to find the reciprocal neighbors quickly.
4. Design and implement an imputation module. This module will incorporate various methods of imputation and must allow incorporation of other methods easily.

References:
[1] Springer, Information Fusion in Data Mining (2003)
[2] Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons
[3] T. Marwala. (2009) Computational Intelligence for Missing Data Imputation, Estimation, and Management Knowledge Optimization Techniques. Information Science Reference

Minimal Requirements & Previous Knowledge:

Basic knowledge on R project language
Good background in mathematics and statistics
Master course-lever knowledge of data mining, machine learning techniques.
Advanced reading and writing english language skills

Other comments:

Location and Date: 23/02/2011 Barcelona,

To the Academic Commission of the Master in Artificial Intelligence (CAIMIA)