

Using CMAT for Data Mining of Large Data Sets

W. M. Hartmann, D-69117 Heidelberg, www.cmat.pair.com/cmat

Abstract In Data Mining we are confronted with either many observations (1), many variables (2), or many of boths (3). In this presentation we discuss some newer strategies for tackling (1) and (2), for example,

1. the possible use the transposed data set and the Sherman-Morrison-Woodbury formula;
2. the case of many observations (chunking, sampling),
3. the case of many variables: use of some variable selection and dimension reduction methods.

The strategies are discussed in connection with known statistical analysis methods, like principal components, variable clustering, SVM, SCAD, and MDS.

References

- [1] Fung, G. & Mangasarian, O.L. (2003), “A Feature Selection Newton Method for Support Vector Machine Classification”, *Computational Optimization and Applications*, 1-18.
- [2] Hastie, T., & Tibshirani, R., (2004), “Efficient quadratic regularization for expression arrays”, *Biostatistics*, **5**, 329-340.
- [3] Mangasarian, O.L. & Thompson, M.E. (2006), “Massive data classification via unconstrained support vector machines”, *Journal of Optimization Theory and Applications*. Technical Report 06-07, Data Mining Institute, University of Wisconsin, Madison, Wisconsin.
- [4] Mangasarian, O.L. & Thompson, M.E. (2006), “Chunking for massive nonlinear kernel classification”, Technical Report 06-07, Data Mining Institute, University of Wisconsin, Madison, Wisconsin.

Keywords

Large Data Sets, SVD, Sherman-Morrison-Woodbury, sampling and chunking