Abstract

In Data mining and Knowledge Discovery hidden and valuable knowledge from the data sources is discovered. The traditional algorithms used for knowledge discovery are bottlenecked due to wide range of data sources availability. Class imbalance is a one of the problem arises due to data source which provide unequal class i.e. examples of one class in a training data set vastly outnumber examples of the other class(es). This paper proposes a method belonging to under sampling approach which uses OPTICS one of the best visualization clustering technique for handling class imbalance problem. In the proposed approach, further Classification of new data is performed by applying C4.5 algorithm as the base algorithm. The method is optimized by the selection of the most suitable clusters for deletion of the majority dataset based on visualization algorithms. An experimental analysis is carried out over a wide range of highly imbalanced data sets and uses the statistical tests suggested in the specialized literature. The results obtained show that our novel proposal outperforms other classic and recent models in terms of Area under the ROC Curve, F-measure, precision, TP rate and TN rate.
enences

- EnhongChe, Yanggang Lin, HuiXiong, QimingLuo, Haiping Ma, &quot;Exploiting probabilistic topic models to improve text categorization under class imbalance&quot;, Information Processing and Management 47 (2011) 202–214.
- Haibo He, Member, IEEE, and Eduardo A. Garcia, &quot;Learning from Imbalanced Data&quot;, IEEE Transactions on knowledge discovery and engineering , Vol 21, No. 9, September 2009.
- MihaelAnkerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander, &quot;OPTICS: Ordering Points To Identify the Clustering Structure&quot;, Proc. ACM SIGMOD&amp;apos;99 Int. Conf. on Management of Data, Philadelphia PA, 1999.

Index Terms

Computer Science  
Artificial Intelligence
Keywords
Classification  class imbalance  CIL-OP