ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Critical Nuggets Identification for Classification Task

V.Priya[1]

Dept. of CSE, Sri Krishna College of Engineering and Technology, Coimbatore, Tamilnadu, India[1]

ABSTRACT: Pattern detection and outliers has emerged as an important area of work in the field of data mining. Critical nuggets are small collections of records or instances that contain domain specific important information for classification. The system identifies the critical nuggets to measure the criticality of data instances. Critical nuggets are identified using CR_{score} which helps in improving classification accuracies in real-world data sets. The CR_{score} values are stored in histogram and highest score values are identified. Using the highest CR_{score} the critical nugget is identified in the data sets. It improves the accuracy of classification. Class imbalance problems have drawn growing interest recently because of their classification difficulty caused by the imbalanced class distributions.

I. INTRODUCTION

Data Mining is the process of extracting knowledge hidden from large volume of raw data. The importance of collecting data that reflect business or scientific activities to achieve competitive advantage is widely recognized now. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining automates the process of finding relationships and patterns in raw data and delivers that can be either utilized in an automated decision support system. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important to develop powerful means for analysis and perhaps interpretation of such data and extraction of interesting knowledge that could help in decision making. In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind if information repository.

Classification task

Classification is a data mining function that assigns items in a collection to target categories of classes. A Classification task begins with a data set in which the class assignments are known .In classification task, the main goal is to derive an accurate representative data model that can correctly classify new test data instances. The accuracy of the classification model can be affected by the presence of outliers in a data set and the inability to correctly classify data records near the boundary.

Classification accuracy

Accuracy is used to measure the classification model. It measures how a binary classification test correctly identifies a condition. A measurement system can be accurate but not precise, precise but not accurate, neither, or both, The result would be a consistent yet inaccurate string of results from the flawed experiment. Eliminating the systematic error improves accuracy but does not change precision.

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

On the other hand, precision or positive predictive value is defined as the proportion of the true positives against all the positive results

$$\text{precision} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{false positives}}$$

Outlier detection

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. The non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, and surprises. The importance of anomaly detection is due to the fact that anomalies in data translate to significant (and often critical) actionable information in a wide variety of application domains. An outlier is defined as a data point which is very different from the rest of the data based on some measure. Such point are often contains useful information on abnormal behavior of the system described by the data.

Problem Definition

Detecting patterns and outliers has emerged as an important area of work in the field of data mining. Existing work not much focused on finding critical nuggets of information that may be hidden in data sets. These nuggets of information may not always be detected by pattern mining methods or by distance-based outlier detection methods as nuggets may not conform to a specific pattern and may not be outliers.

II. EXISTING SYSTEM

In existing system, patterns detection and outliers has emerged as an important area of work in the field of data mining. Outliers are abnormal objects that deviate from other normal objects. Finding the outliers in several applications such as credit card fraud detection and identifying network intrusions help in identifying the outliers. The system provides the idea of finding the outliers detection based on their difference in the object distance. In Classification task, by detecting the outlier in the data set provides less accuracy result obtained. The mining of outliers provides to indentify records that are different from rest of the data sets. But, the existing work not much focuses on finding critical nuggets of information in the data sets. The nuggets of information are not been detected by outlier detection.

Disadvantage of the existing system:

- Less classification accuracy.
- Do not overcome the class imbalance problem.

III. PROPOSED SYSTEM

In proposed system, Critical nuggets are small collections of records or instances that contain important information. The system provides the critical nuggets to measure the criticality of data instances. Critical nuggets are identified using the CRscore calculation and important instance are identified using the score value. It also helps in improving classification accuracies in real-world data sets. Class imbalance problems have drawn growing interest recently because of their classification difficulty caused by the imbalanced class distributions. Applying, a preprocessing step in order to balance the class distribution for imbalanced dataset problem, so the accuracy of the critical nugget can be improved by the process. Further multi class problem is also overcome in the proposed system.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Advantage of proposed system:

- Improves the classification accuracy
- Class imbalance is overcome by fine turning.

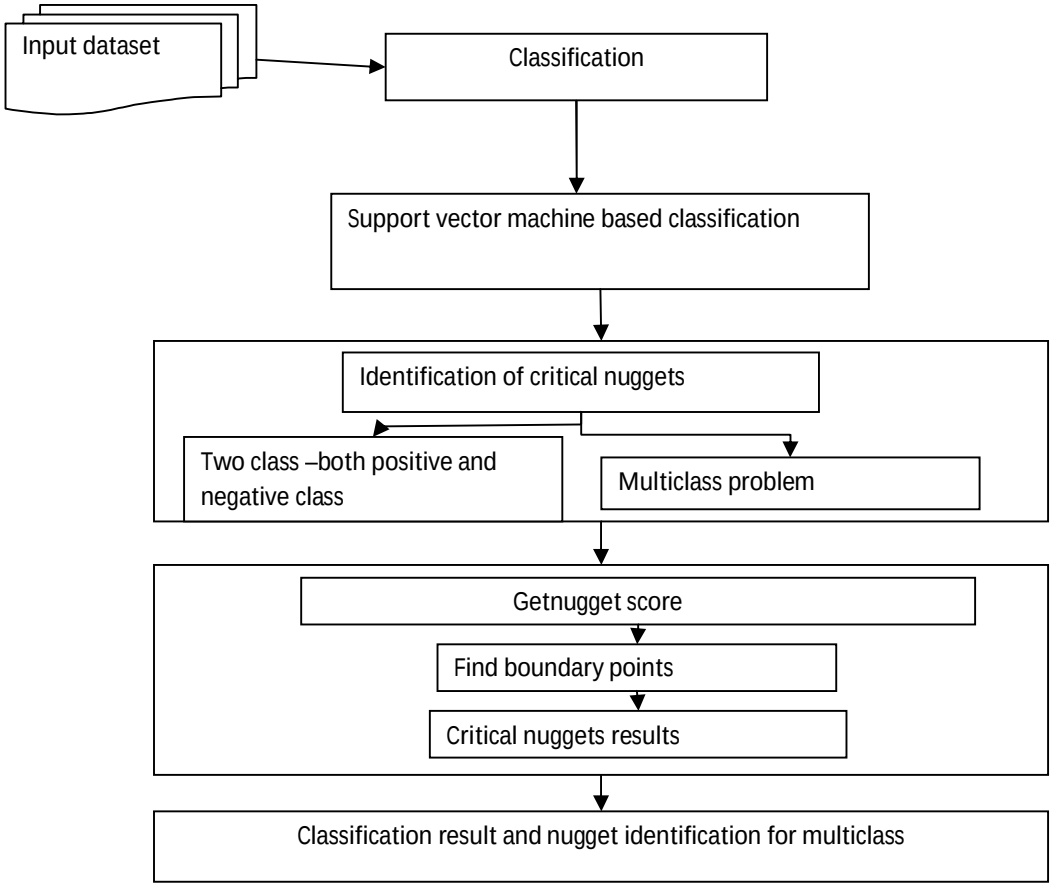

Figure 1.1 Process involved in critical nugget

IV. CONCLUSION

Initially classification training file is formed and a new calculation CRscore is introduced for measuring the criticality of a subset or nugget. For each column in the attributes, corresponding changes are made to find the critical data. Then, additional concepts are used to resolve conflicting scores when they occur. The critically helps in finding the critical information during the classification task. Finally, the critical information are identified using the CR score value and the values are added to the histogram and the reduced instances are used training. The critical instance is used in the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

classification task for predicting the classification accuracy. This helps in improving the classification accuracy in the given datasets.

REFERENCES

[1]E.M. Knorr, R.T. Ng, and V. Tucakov, "Distance-Based Outliers:Algorithms and Applications," VLDB J., vol. 8, no. 3/4, pp. 237-253, 2000.
[2] M.M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," SIGMOD Record,vol. 29, no. 2, pp. 93-104, 2000.
[3] Charu C.Aggarwal and Philip S.Yu,"Outlier Dection for High Dimensional Data" Data Mining and Knowledge Discovery, vol. 11, no. 6,2001
[4] N. Panda, E.Y. Chang, and G. Wu, "Concept Boundary Detection for Speeding Up SVMs," Proc. 23rd Int'l Conf. Machine Learnin g(ICML), W.W. Cohen and A. Moore eds., vol. 148, pp. 681-688,2006.
[5] Y. Tao, X. Xiao, and S. Zhou, "Mining Distance-Based Outliers from Large Databases in Any Metric Space," Proc. 12th ACMSIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD),T. Eliassi-Rad, L.H. Ungar, M. Craven, and D. Gunopulos, eds.,pp. 394-403, 2006.
[6] A. Ghoting, S. Parthasarathy, and M.E. Otey, "Fast Mining of Distance-Based Outliers in High-Dimensional Datasets," DataMining and Knowledge Discovery, vol. 16, no. 3, pp. 349-364,2008.
[7] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection:A Survey," ACM Computing Survey, vol. 41, no. 3, 2009.
[8] R.A. Weekley, R.K. Goodrich, and L.B. Cornman, "An Algorithm for Classification and Outlier Detection of Time-Series Data,"J. Atmospheric and Oceanic Technology, vol. 27, no. 1, pp. 94-107, Jan.2010.