Imbalanced Classification in Predictive data mining using GSVM Algorithm

N.Sathyabhama
Assistant Professor, Department of MCA, SNS College of Technology, Coimbatore

Dr.S.V.Saravanan
Principal-CKEC Coimbatore

V.Srimathi
Assistant Professor ,Department of MCA, SNS College of Technology, Coimbatore

Abstract- The paper addresses some theoretical and sensible aspects of information mining, specializing in prophetical data processing, wherever 2 central forms of prediction issues are discussed: classification and regression. additional accent is formed on prophetical data processing, wherever the time-stamped information greatly increase the size and complexness of downside resolution. the most goal is thru process of information (records from the past) to explain the underlying dynamics of the complicated systems and predict its future. Traditional classification algorithms may be restricted in their performance on extremely unbalanced datasets. a preferred stream of labor for countering the matter of sophistication imbalance has been application of a sundry of sampling ways. During this work, we have a tendency to specialize in the matter of sophistication imbalance. We have a tendency to incorporate completely different "rebalance" heuristics.

Keywords – Data Mining, Imbalanced data sets, Logistic regression, Predictive data mining, Sampling

I. INTRODUCTION

Solution of the issues regarding water resources and surroundings nowadays depends on sizable amount of information sources and knowledge corpuses. Several relevant sources of information, structured observations and scientific info associated with water resources and environmental processes presently exist, variable in each size and scope. the massive potentials within the existing information banks have to be compelled to be explored so as to rework these data/observables into valuable engineering info and data. The key to those potentials will be found in data processing as a replacement up corning filed in hydro scientific discipline.

Mining extremely unbalanced datasets, significantly in a very value sensitive surroundings, is among the leading challenges for data discovery and data processing. the category imbalance drawback arises once the category of interest is relatively rare as compared to the opposite class(es). while not loss of generality we'll assume that the positive category (or category of interest) is that the minority category, and therefore the negative category is that the majority category. numerous applications demonstrate this characteristic of high category imbalance, like bioinformatics, e-business, info security, to national security. as an example, within the medical domain the unwellness could also be That is, either the minority category is oversampled or majority category is below sampled or some combination of the 2 is deployed.

.

II. DATA MINING

Data mining can be defined as a process of discovering new, interesting knowledge, such as patterns, associations, rules, changes, anomalies and significant structures from large amounts of data stored in data banks and other information repositories. It is currently regarded as the key element of a much more elaborate process called Knowledge Discovery in Databases (KDD). In general, a knowledge discovery process consists of an iterative

sequence of the following steps (see *Figure 1*):

1. Data selection, where data relevant to analysis task are retrieved from database;

2. *Data cleaning*, which handles noisy, erogenous, missing or irrelevant data;

3. *Data integration (enrichment)*, where multiple heterogeneous data may be integrated into one;

4. *Data transformation (coding)*, where data are transformed or consolidated into forms appropriate for different mining algorithms;

5. *Data mining*, which is an essential process where intelligent methods are applied in order to extract hidden and valuable knowledge from data;

6. *Knowledge representation*, where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

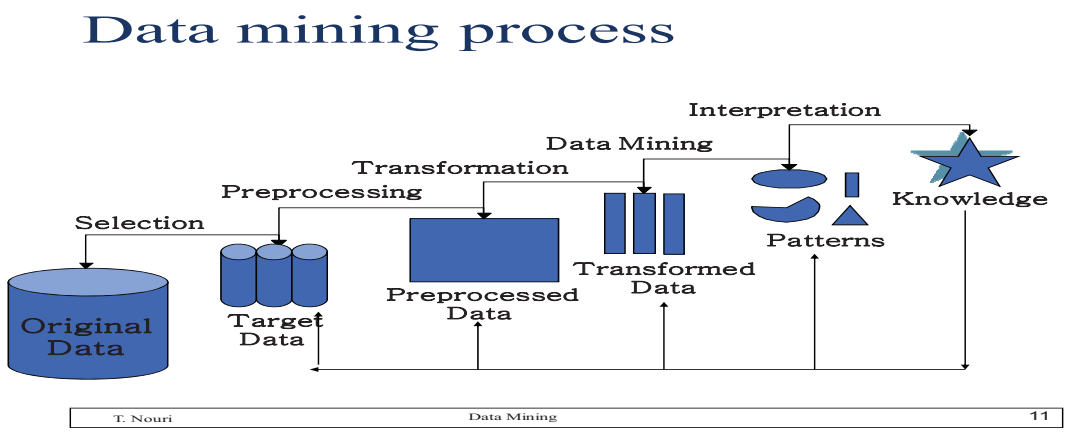

Figure: 1

Data mining is based on the results achieved in database systems, statistics, machine learning, statistical learning theory, chaos theory, pattern recognition, neural networks, probabilistic graph theory, fuzzy logic and genetic algorithms. A large set of data analysis methods have been developed in statistics over many years of studies. Machine learning and statistical learning theory have contributed significantly to classification and induction problems. Neural networks have shown their effectiveness in classification, prediction, and clustering analysis tasks. One can say that there is no one specific technique that characterizes data mining. Any technique that helps to extract more out of the data sets in an autonomous and intelligent way may be classified as a data mining technique. Therefore data mining techniques form a quite heterogeneous group.

In general, data mining tasks can be classified into two categories:

Description: finding human-interpretable patterns, associations or correlations describing the data.

Prediction: constructing one or more sets of data models (rule set, decision tree, neural nets, and support vectors), performing inference on the available set of data, and attempting to predict the behavior of new data sets. The distinction between description and prediction is not very sharp. Predictive models can also be descriptive (to the degree that they are understandable), and descriptive models can be used for prediction. To achieve these goals, the categories of prediction as well as description are associated with the five basic operations.

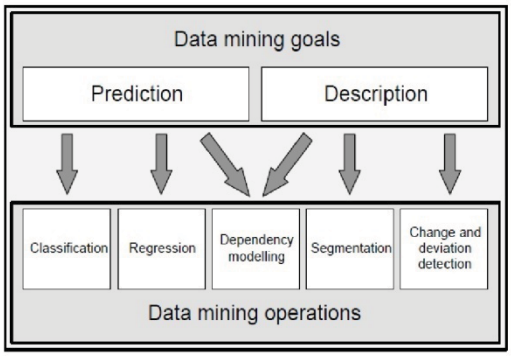

Figure 2. The connection between data mining goals and operations

The two central types of engineering prediction problems are *classification* and *regression*. Samples/observables of past experience with known attributes (features) are examined and generalized to future cases. Classification is closely coupled with clustering which is to identify clusters embedded in the multi-dimensional data space, where a cluster is a collection of data objects (groups of data) that are "similar" to one another. Similarity usually is expressed as different distance functions. Various approaches have been proposed in the literature for developing classifiers by means of clustering, which can be summarized as:

(i) Iterative clustering, (ii) Agglomerative hierarchical clustering and (iii) Divisive hierarchical clustering.

From a perspective of data mining, classification and clustering algorithms that employ unsupervised learning receive greater attention. The reason for this lies in the fact that in most of engineering classification problems the set of possible classes is not known *a priori*. The goal is to find the classes themselves from a given set of "unclassified" objects/observables which may lead to discovery of previously unknown structure, because in natural systems (such as water and environment related systems) there are usually many relevant attributes describing each object - large number of dimensions.

However, we wish to emphasize that unsupervised classification task should usually come together with the background knowledge provided from the domain experts.

The problem of regression is very similar to the problem of classification. It is usually described as a process of induction of the data model of the system (using some machine learning algorithm) that will be capable of predicting responses of the system that have yet to be observed. For regression the response of the system is usually a real value, while for classification is the class label(s). Time series prediction is a specialized type of regression (or occasionally classification) problem, where measurements/observables are taken over time for the same features.

From a predictive data-mining perspective, the time-stamped data greatly increase the dimensions of problem solving in a completely different direction. Instead of cases with one measured value for each feature, cases have the same feature measured at different time. To overcome this problem, raw time-dependent data are usually transformed for predictive data mining into lesser dimensional data space using transformations such as Vector Quantization and state-space methods (Tsonis, 1992) or simple averaging and re-sampling methods are applied.

The main goal of this work is to demonstrate the applicability of some predictive data mining techniques for classification and regression engineering problems.

III. Metrics For Measuring Predictive Model Performance

The different types of errors and hits performed by a classifier can be summarized in a matrix. Table 1 illustrates a matrix for a two-class problem, with classes labeled positive and negative.

Table 1. Different types of errors and hits for a two classes problem.

Actual ▼ Predicted ▸	*Positive*	*Negative*
Positive	True positive (a)	False positive (b)
Negative	False negative (c)	True negative (d)

| Negative | False negative (c) | True negative (d) |

Error rate (E) = (c+b) /(a+b+c+d)

Accuracy (Acc) = (a+d) /(a+b+c+d) = 1 - E.

For instance, it is straightforward to create a classifier having 99% accuracy (or 1% error rate) if the data set has a majority class with 99% of the total number of cases, by simply labeling every new case as belonging to the majority class.

Following is the proposed the metrics using Table 1.

– False negative rate: FN = b / (a+b) is the percentage of positive cases misclassified as belonging to the negative class;

– False positive rate: FP = c / (c+d) is the percentage of negative cases misclassified as belonging to the positive class;

– True negative rate: TN = d / (c+d) = 1 - FP is the percentage of negative cases correctly classified as belonging to the negative class;

– True positive rate: TP = a / (a+b) = 1 - FN is the percentage of positive cases correctly classified as belonging to the positive class;

These four class performance measures have the advantage of being independent of class costs and prior probabilities. It is obvious that the main objective of a classifier is to minimize the false positive and negative rates or, similarly, to maximize the true negative and positive rates. Unfortunately, for most "real world" applications, there is a tradeoff between FN and FP and, similarly, between TN and TP.

IV. Metrics For Imbalanced Classification

Many metrics have been used for effectiveness evaluation on imbalanced classification. All of them are based on the matrix as shown at Table I. With highly skewed data distribution, the overall accuracy metric at (1) is not sufficient any more. For example, a naive classifier that predicts all samples as negative has high accuracy. However, it is totally useless to detect rare positive samples. To deal with class imbalance, two kinds of metrics have been proposed.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$f - value = \frac{(1 + \beta_2) \times precision \times recall}{\beta_2 \times recall + precision} \quad (3)$$

To get optimal balanced classification ability, sensitivity at (2) and specificity at (3) are usually adopted to monitor classification performance on two classes separately. Notice that sensitivity is also called true positive rate or positive class accuracy, while specificity called true negative rate or negative class accuracy. Based on these two metrics, G-Mean was proposed at (4), which is the geometric mean of sensitivity.

V. GSVM Algorithm

Granular computing represents data within the type of some aggregates (called data granules) like subsets, subspaces, classes, or clusters of a universe. It then solves the targeted drawback in every data grain. There are 2 principles in granular computing. the primary principle is divide-and-conquer to separate an enormous drawback into a sequence of grains (granule split); The second principle is information improvement to outline the appropriate size for one granule to grasp the matter at hand while not obtaining buried in unessential details (granule shrink). As hostile ancient data-oriented numeric computing, granular computing is knowledge-oriented. By embedding previous information or previous assumptions into the granulation method for information modeling, higher classification may be obtained. A granular computing-based learning framework known as Granular Support Vector Machines.

For a extremely unbalanced dataset, there could also be several redundant or howling negative samples. Random under sampling could be a common under sampling approach for rebalancing the information set to attain higher data distribution. However, random under sampling suffers from data loss.

The well-known reality regarding SVM - solely SVs are necessary and alternative samples may be safely removed while not moving classification. This reality motivates U.S. to explore the likelihood to utilize SVM for information cleaning/ under sampling. However, thanks to extremely inclined information distribution, the SVM shapely on the initial coaching dataset is vulnerable to classify each sample to be negative. As a result, one SVM cannot guarantee to extract all informative samples as SVs. associate degree aggregation operation is then dead to by selection combination the samples within the negative data granules with all positive samples to finish the under sampling method. Finally, associate degree SVM is shapely on the collective dataset for classification.

VI. Conclusion

In this work we have a tendency to mentioned and incontestable a helpful approach of the prophetical data processing techniques that specialize in 2 styles of engineering drawback finding classification and regression. once handling great deal of knowledge and once categories got to be discovered. Its straightforward nature and therefore the applied mathematics background create this approach powerful data processing tool, particularly once combined

with the domain data.

The Granular Support Vector Machines formula implements a guided repetitive under sampling strategy to "rebalance" the dataset at hand. it's effective as a result of extraction of informative samples that area unit essential for classification and elimination of an oversized quantity of redundant or maybe strident samples

REFERENCES

[1] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," Intelligent Data Analysis, vol.6, no. 5, pp. 429–449,2002.

[2] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data set" SIGKDD Explorations, vol.6, no.1, pp 1-6,2004.

[3] Provost, F., Jensen, D., and Oates, T. Efficient Progressive Sampling. Proceedings of the Fifth International Conference on Knowledge discovery and Data Mining, 23-32, ACM Press 1999

[4] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations, 6(1), 2004.

[5] M. Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, 2003.

[6] N. V. Chawla. Editorial: Learning from Imbalanced Datasets. *SIGKDD Explorations*, 6(1), 2004.

[7] L. Ghouti, A. Bouridane, M.K. Ibrahim, and S. Boussakta, "Digital image watermarking using balanced multiwavelets", *IEEE Trans. Signal Process*., 2006, Vol. 54, No. 4, pp. 1519-1536.

[8] P. Tay and J. Havlicek, "Image Watermarking Using Wavelets", in *Proceedings of the 2002 IEEE*, pp. II.258 – II.261, 2002.

[9] P. Kumswat, Ki. Attakitmongcol and A. Striaew, "A New Approach for Optimization in Image Watermarking by Using Genetic Algorithms", *IEEE Transactions on Signal Processing*, Vol. 53, No. 12, pp. 4707-4719, December, 2005.

[10] H. Daren, L. Jifuen,H. Jiwu, and L. Hongmei, "A DWT-Based Image Watermarking Algorithm", in *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 429-432, 2001.

[11] C. Hsu and J. Wu, "Multi-resolution Watermarking for Digital Images", *IEEE Transactions on Circuits and Systems- II*, Vol. 45, No. 8, pp. 1097-1101, August 1998.

[12] R. Mehul, "Discrete Wavelet Transform Based Multiple Watermarking Scheme", in *Proceedings of the 2003 IEEE TENCON*, pp. 935-938, 2003.