DOMAIN KNOWLEDGE IDENTIFICATION AND CLASSIFICATION FOR DISEASE DIAGNOSIS

Jothilakshmi.M, Asst.Prof Anuvelavan.S/CSE
University College Of Engineering, Trichirappalli

Abstract

The collection contains minimum amount of data that contains important information which is used to perform classification for the particular domain. These records collectively removed from the dataset called criticality measure and the percentage of removing records are calculated by using criticality score (CR score) value. The class boundary is identified by using find boundary algorithm and small set of data used for classification called critical nuggets are identified by using find critical nuggets algorithm. Find critical nuggets algorithm is divided into two phases for two classes. The identification and classification method is enhanced to support multi class environment and the system need to adapt mixed attribute values. Such attribute types are binary, continuous, category. Centroid relationship data considered as relevant class and to reduce the difficulty of nuggets detection, boundary approximation algorithm is developed.

IndexTerms–Datamining,Classification,Outlier,Nuggets identification

1.Introduction

Data mining is an essential step in the process of knowledge discovery in databases in which intelligent methods are applied in order to extract patterns. Breast cancer has become the primary reason of death in women in developed countries. The most effective way to reduce breast cancer deaths is to detect it earlier. Early diagnosis needs an accurate and reliable diagnosis procedure that can be used by physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy. The objective of these predictions is to assign patients to one of the two group either a “benign” that is noncancerous or a “malignant” that is cancerous. The prognosis problem is the long-term care for the disease for patients whose cancer has been surgically removed.

Predicting the outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications. The use of computers with automated tools, large volumes of medical data are being collected and made available to the medical research groups. As a result, data mining techniques has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict the outcome of a disease using the historical datasets. The objective of this study is to summarize various review and technical articles on diagnosis and prognosis of breast cancer.

Mining for outliers in data is an important research field with many applications in credit card fraud detection, discovery of criminal activities in electronic commerce, and network intrusion detection. Outlier detection approaches focus on discovering patterns that occur infrequently in the data, as opposed to many traditional data mining techniques, such as association analysis or frequent itemset mining, that attempt to find patterns that occur frequently in the data. One of the most

ISSN: 2278 – 1323

widely accepted definitions of an outlier pattern is provided by Hawkins: "An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism".

2. Related Work

In classification problems, 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.

Considering the first case of outliers are critical nuggets different from outliers and can existing approaches in outlier detection help in finding critical nuggets? Critical nuggets in certain cases may involve outliers, but this may not always be true. In the example of the previous section, cells in tumors may not show anomalous behavior on an individual basis but collectively, such cells may contain critical pieces of information. In [1], the authors note that the performance of a distance-based outlier detection method "greatly relies on a distance measure, defined between a pair of data instances, which can effectively distinguish between normal and anomalous instances. Defining distance measures between instances can be challenging when the data are complex."

Moreover, critical nuggets that belong to a data set may not be at a great "distance" from the other "normal" points, and may end up being classified as "normal." For a comprehensive survey on outlier detection methods, please refer to the extensive survey in [1]. In the field of distance-based outlier detection, researchers have focused on proposing algorithms that reduce the time complexity $O(n^2)$ of calculating distances [2], [9], [8], and [3]. Work has also been done on density-based outlier detection such as [4] where outliers are defined as objects that show anomalous trends with respect to their local neighborhoods and tend to lie in a less dense area with respect to a more dense local neighborhood. In [5], the concept of density-based detection is extended to cluster-based outlier detection where the approach does not only find single point outliers but instead clusters of outliers. Intuitively, cluster-based outlier methods may not necessarily lead to identifying critical areas, which do not lie at a great
"distance" from the rest of the points.

With this differentiation between critical nuggets and outliers, can critical nuggets be found among data records near the boundary? One can utilize an intuition that is motivated by two commonly occurring scenarios in classification algorithms:

- Points near the boundary, in general, are critical. The deciding factor for most classification algorithms is how accurately the algorithm classifies the points near the class boundaries (see also [6]). The points that are far from the class boundaries are the

  "slamdunk," easy cases, where the impact of misclassification is pretty minimal. However, the points near the class boundaries are more susceptible to misclassification.
- Certain boundary features can be critical. Second, as a corollary to the first scenario, there are certain regions along the boundary where the problem of classification becomes more difficult, as compared to less problematic boundary points.

In summary, using the first scenario, the search for critical nuggets is narrowed to a region near the boundary separating the classes. On the basis of the second scenario, where certain boundary features are more complex than others,

ISSN: 2278 – 1323

the criticality metric (the $CR_{score)}$ has been defined in such a way that it yields higher scores for sets of data records that lie near complex boundary features. In other words, the greater the complexity of a boundary feature, the higher the probability of misclassification becomes, resulting in higher scores being assigned for points near that complex boundary.

1. Classification and Critical Nuggets

The study of observing data that differs at most from other data called outlier. These data are not used to perform classification, but it may alert the user of providing false rate of the process. Here introduce new concept called nuggets, comes under the category of outlier. Nuggets may be an outlier, but no assurance of having the nuggets under the outlier category. The small amount of data which is used to perform classification called critical nuggets. The group of data to be removed from the relevant class and moved to some other class called criticality. The classification model taken as M1 and from that class model, subset of data to be removed (called N) and it form new class called M2 based on some conditions. Consider the below term for the criticality,

$$M2 = M1 - N$$

M2 – New class
M1 – Existing class
N – Collection of data to be removed

The idea is to find the number of attributes is very sensitive while performing small changes in a class. These attributes are moved from relevant class to some other class when the changes are performed. Find the number of attributes moving from one to another class call criticality score value (CRscore). For example out of four direction, three direction data are mover from one to another class while doing some changes.

Criticality score value is calculated as 3/4=75. So 75% of data are moved from existing class new class. In large data set get nugget score algorithm is used to find the critical attributes which is having the chance of moving from one class to some new class. This algorithm using rotation method algorithm to find criticality score value in large and high dimensional database.

Critical nuggets are identified by using criticality score value. The class boundary is identified to differentiate one class from another. Get nugget score revised algorithm used to list the score value of attribute. Threshold value is identified by using class boundary. If the data nearer to threshold value, this is having chance of moving from one class to another. Get nugget score revised algorithm works in two phases. First phase working in relevant class and second phase works in moved class. Find critical nuggets algorithm used to detect critical nuggets to improve the classification accuracy. This algorithm works in two phases, each phase working in any one of the two classes. Steps of critical nuggets identification is below,

- Identify the approximate boundary
- The neighborhood data are considered around the boundary set.
- Based on CR_{score} value, critical nuggets are identified.

Get nuggets score algorithm is used to identify the criticality score value. Criticality score value is used to detect critical nuggets by using get nuggets score algorithm. In large dataset it is difficult to identify the small nuggets by using criticality score value. To overcome this problem class boundary algorithm is to be developed. In this method boundary is identified which is used to separate two classes. The idea is points near to the boundary having the chance of moving from one to next class which is identified as critical nuggets.

ISSN: 2278 – 1323

Find critical nuggets algorithm works in two phases. One is to identify the critical nuggets in relevant class and remaining is new class or moving class. The goal is to find critical nuggets in one class at a time. Because the boundary is created by using the distance value R and based on center of point, nuggets are identified. Identify the attributes which belong to the class and it is moving to another class when the attribute values are changed. Critical nuggets are having duality property. The critical nuggets belong to one class which is very closer to the critical nuggets of another class is called duality of critical nuggets. This property to be verified while using find critical nuggets algorithm for classification purpose.

The system is used to identify the critical nuggets based on criticality score value analysis and in large data area; the critical nuggets are identified by using the class boundary to improve the classification accuracy. The center of class data is used to identify the nuggets for the relevant class. The existing system has following disadvantages.

- Boundary estimation is complex
- Not supports mixed attribute environment
- Classification is limited with two class levels.

4.Attribute Dependant Classification Process

4.1. Criticality Scores

Consider a training data set T_r with m data instances, each instance having n attributes denoted as A_j ($j\varepsilon\{1, 2, \ldots, n\}$). The underlying assumption is that all attributes are numeric and not categorical. From T_r, form a neighborhood N, by choosing a data instance D_i as a center and finding a group of points that belong to the same class as D_i and lying within a distance R from D_i. For simplicity, let us say that the neighborhood N is comprised of d data instances [7]. The selection of parameters R and D_i used in forming a neighborhood N. First, a classification model M_0 is generated by applying a classification algorithm C to the training data set T_r. Using the classification model M_0, one can predict the class labels for the different data instances in question. For the d instances in neighborhood N, consider an attribute A_j. Also, for the d instances, the attribute A_j can be increased or decreased in magnitude. A parameter denoted by δ_j is used for this and δ_j varies for different attributes in neighborhood N. After increasing A_j by an extent δ_j for just the d instances, the classification model M_0 for the new class labels for the d instances is queried. The average number of data instances that have switched classes in neighborhood N is computed and is denoted as w_j^-. If all the data instances in N switch classes, then one can infer that N is very sensitive to changes with respect to attribute A_j. The same test is applied on N by decreasing A_j by the same extent δ_j and find w_j^+ by querying the classification model M_0 for the new class labels. For the attribute A_j, the average of w_j^- and w_j^+ is computed to get w_j. Repeating this process for all n attributes, the average of the w_j scores is computed as the CR_{score} for the neighborhood N. Formally, the critical score is defined as

$$CR_{score} = \frac{\sum_{j=1}^{n} w_j}{n} \quad - \quad (1)$$

where: $w_j = \frac{w_j^+, w_j^-}{2}$, $w_j^+ = \frac{d_j^+}{d}$, and $w_j^- = \frac{d_j^+}{d}$.

Using the description on how the CR_{score} is calculated, the algorithm GetNuggetScore is developed. The computational complexity of the algorithm is derived as follows: Deriving the model M_0 is dependent on the complexity of the chosen classification algorithm (C). The complexity of the classification algorithm is

denoted as t(C). Each attribute A_j is analyzed by checking if increasing or decreasing the values of the attributes by an extent δ_j, switches the class label. Hence, for each attribute, the model M_0 is queried twice. There are d data instances in N and, thus, for each attribute there are $2 \times d$ queries. Since there are n attributes, the complexity of the for-loop is O(dn). When $d \ll n$, the complexity of the for-loop becomes $\approx O(n)$. The total complexity of the algorithm is O(t(C) + dn).

The FindCriticalNuggets algorithm works in two phases. In each phase it identifies critical nuggets for each one of the two classes. Using the reduced boundary set for each class, the data instances in the boundary set are considered one at a time. Each data instance in the boundary set is considered as a center for a neighborhood. A neighborhood is formed by finding all points that belong to the same class and lie within a distance R from the center point. One class at a time is considered because the goal is to find critical nuggets that belong to one class but switch to the other class when their attribute values are perturbed. If there are $|B^+|$ data instances in the boundary set that belong to the same class (say "+"), one can form $|B^+|$ neighborhoods by considering each instance in B^+ as a center. For each of the $|B^+|$ neighborhoods, the CR_{score} is computed. The scores are then ranked and the higher scores are used to identify the critical nuggets in T_r^+. The other class is considered. Hence, $|B^+|$ neighborhoods are then considered to compute the CR_{score} values, which in turn are sorted and ranked to identify critical nuggets in T_r^+.

4.2. Classification Process

Classification algorithms are usually judged based on the accuracy of their predictions. If the predictions include a minimum number of false positives and false negatives, the accuracy of an algorithm is rated as high. During the experimental stage with various data sets, tests were conducted to see if critical nuggets could help improve the classification accuracy. The identified nuggets were used in deriving additional small scale classification models. For each class, an additional classification model is built / trained by first deriving a new data set, which is a subset of the original training data set. The new data set was derived by relabeling a subset of the original data records into two new classes as follows:

- Data records that belong to the top k critical nuggets become a part of one class.
- Data records that are near the top k "+" class critical nuggets but NOT belonging to the set of "+" class critical nuggets are labeled as another class.

5. Domain Knowledge Identification and Classification

The critical nuggets identification and classification scheme is improved to support multiple classes. The system can be adopted to handle mixed attribute data values. The boundary approximation algorithm is enhanced to reduce the detection complexity. Post processing operations are tuned to identify classes for multiple category data environment.

The critical nuggets based classification system is designed to classify multi-class data values. Multi attribute data analysis mechanism is applied to handle all attribute types. Classification is performed with the support of critical nuggets extracted from the learning process. The system is divided into five major modules. They are data preprocess, nuggets identification, class boundary analysis, classification on bi-class data and classification on multi-class data.

The data preprocess module is designed to perform cleaning operations. Nugget identification

module is designed to fetch critical nuggets from transactions. Class boundary analysis module is used to identify the threshold for classes. Two level class label assignment process is performed under the classification on bi-class data. Multi level class label assignment process is performed under the classification on multi-class data.

5.1. Data Preprocess

Lung cancer data values are collected and analyzed in the data preprocess. Noisy data elements are corrected with suitable values. Aggregation based data substitution mechanism is used to assign values for missing elements. Learning and testing data values are partitioned in the preprocess.

5.2. Nuggets Identification

Nugget identification process is performed on the labeled transactions. Criticality score is estimated for the attributes and transactions with class information. Nugget score values are verified with associated class information. Critical nuggets are identified for each class levels.

5.4. Classification On Bi-class Data

The nugget based classification algorithm is designed to detect two class levels only. The systems select nuggets for two class levels from the labeled transactions. The unlabeled transactions are compared with the nuggets associated with the classes. Similarity analysis is performed between the nuggets and unlabeled transactions for the class assignment process.

5.5. Classification On Multi-class Data

The nugget based classification scheme is tunes to detect multiple class labels. Class boundary identification is also enhanced to support multiclass environment. Multi attribute based classification is performed on the binary, categorical and continuous attributes. Nugget similarity analysis is applied for each class levels.

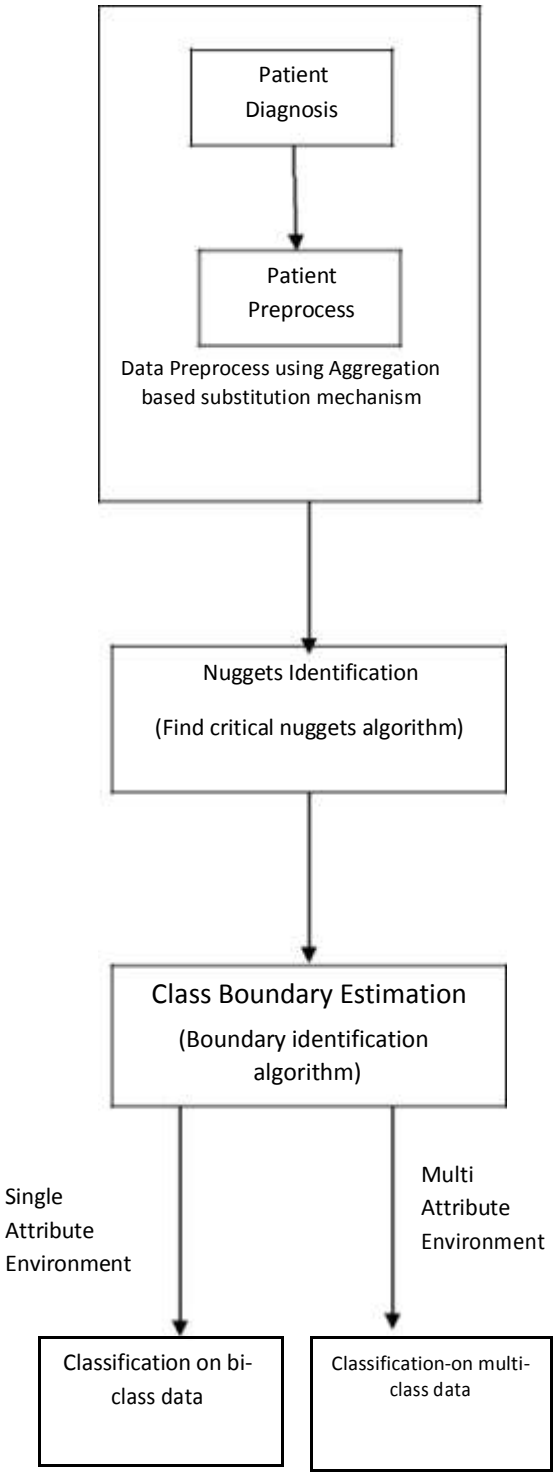

Figure No: 5.1. Domain Knowledge

6. Conclusion

Classification techniques are used to identify the transaction label. Critical nuggets are used to represent the domain knowledge of the data collection. Classification accuracy is improved with critical nuggets and class boundary algorithm. The system is enhanced to support multiple class and multi attribute environment. False positive and false negative errors are reduced in the classification process. Classification accuracy is improved by the nuggets based classification scheme. The system reduces the Computational complexity. The system supports mixed attribute data for classification process.

REFERENCES

[1] V. Chandola and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Survey, vol. 41, no. 3, article 15, 2009.

[2] S.D. Bay and M. Schwabacher, "Mining Distance-Based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule," Proc. Ninth ACM SIGKDD Int'l Conf.(KDD), 2003.

[3] A. Ghoting and M.E. Otey, "Fast Mining of Distance-Based Outliers in High-Dimensional Datasets," Data Mining and Knowledge Discovery, 2008.

[4] M.M. Breunig and J. Sander, "LOF: Identifying Density-Based Local Outliers," SIGMOD Record, 2000.

[5] L. Duan and J. Lee, "Cluster-Based Outlier Detection," Annals of Operations Research, Apr. 2009.

[6] E. Triantaphyllou, Data Mining and Knowledge Discovery via Logic-Based Methods. Springer, 2010.

[7] David Sathiaraj and Evangelos Triantaphyllou, "On Identifying Critical Nuggets of Information during Classification Tasks", IEEE Transactions On Knowledge and Data Engineering, Vol. 25, No. 6, June 2013.

[8] Y. Tao ,S. Zhou, "Mining Distance-Based Outliers from Large Databases in Any Metric Space," Proc. 12th ACM SIGKDD Int'l Conf. (KDD), 2006.

[9] F. Angiulli and C. Pizzuti, "Outlier Mining in Large High-Dimensional Data Sets," IEEE Trans. Knowledge Data Eng., Feb. 2005.

ISSN: 2278 – 1323