Latent Topic Based Medical Data Classification

Jian-hua Yeh and Shi-yi Kuo

Abstract—This paper discusses the classification process for medical data. In this paper, we use the data from ACM KDDCup 2008 to demonstrate our classification process based on latent topic discovery. In this data set, the target set and outliers are quite different in their nature: target set is only 0.6% size in total, while the outliers consist of 99.4% of the data set. We use this data set as an example to show how we dealt with this extremely biased data set with latent topic discovery and noise reduction techniques. Our experiment faces two major challenge: (1) extremely distributed outliers, and (2) positive samples are far smaller than negative ones. We try to propose a suitable process flow to deal with these issues and get a best AUC result of 0.98.

Keywords—classification, latent topics, outlier adjustment, feature scaling

I. INTRODUCTION

Classification problem is one of the major issues in data mining research fields. A classifier decides which class an unknown data to go according to existing historical data and predefined classes. The classification problems in medical area often classify information based on the result of medical diagnosis or description of medical treatment process such as laboratory experiment results, radioactive photography, and some other processes.

Cancer is one of the major leading causes of death for human beings, among them the breast cancer is also the leading causes of death for women. One of the well-known breast cancer detection examinations is through breast X-ray images. In recent years, with the progress of computer technology, the X-ray images are often stored in digital formats. With these digital images, the cancer classification based on the digital information becomes easier than before. The scientists use extracted features from historical medical images to train a well-designed classifier, then the classifier might be able to correctly classify an unknown image data for screening cancerous patients.

To determine whether a patient is cancerous is a typical one-class classification problem, since the classification focuses on catching the features of cancer images(called target set), and the others are treated as outliers[1]. In this paper, we use the data from ACM KDDCup 2008[2] to demonstrate our classification process based on latent topic discovery. In this data set, the target set and outliers are quite different in their nature: target set is only 0.6% size in total, while the outliers consist of 99.4% of the data set. We use this data set as an example to show how we dealt with this extremely biased data set with latent topic discovery and noise reduction techniques.

II. RELATED WORK

There are typically two ways to separate data in groups appropriately: grouping by their features without indication, called clustering, and forming groups by existing classes, called classification. Clustering is known as unsupervised learning and classification is known as supervised learning. In statistics domain, the supervised learning is often called discrimination which uses correctly classified data to build discrimination rules. There are several evaluation aspects for classification: accuracy, speed, comprehensibility, and time to learn. The KDDCup 2008 contest focuses on accuracy competition, that is, to compete for the best classification accuracy for given data set. As mentioned earlier, the outliers consist of the major part of the data set, which is about 99.4% of original data. So the outlier is a major problem to affect the classification accuracy in KDDCup 2008. There are two types of error caused by outlier:

i. Type I error: the classifier classifies members of target set as outliers.

ii. Type II error: the classifier classifies members of outlier as target set.

These errors seriously affect the accuracy of classification in our problem. In this paper, we will propose a process flow to minimize these errors.

The data provided in KDDCup 2008 is described as follows: every X-ray sample consists of four X-ray images, called MLO and CC, each contains two images. MLO and CC represent images shooting from different angles. So each patient has four X-ray images as source data. Each image will then be described as candidates, that is, suspicious points. Each candidate is then described by several attributes: image-ID, patient-ID, coordinates (x, y), and some other numerical attributes. Finally, the cancerous candidates will be labeled. The numerical attributes are generated by standard image processing algorithms, total 117 of them. In training data set, a lesion-ID is also provided for candidate data but is missing from the test set. From the patient’s perspective, there are 118 cancerous patients and 1,594 normal patients, which generate 102,294 candidates with 117 features for each candidate.

Jian-hua Yeh is with the Department of Computer Science and Information Engineering, Aletheia University, Taiwan, R.O.C. (phone: 886-2-26212121; e-mail: jhyeh@mail.au.edu.tw).

Shi-yi Kuo is with the Department of Computer Science and Information Engineering, Aletheia University, Taiwan, R.O.C. (phone: 886-2-26212121; e-mail: FM970298@smail.au.edu.tw).

International Scholarly and Scientific Research & Innovation 5(5) 2011 476
research focuses aim at topic detection in textual data by using term distribution calculation among the documents. Several important algorithms were developed, including Latent Semantic Analysis (LSA)[4], Probabilistic Latent Semantic Analysis (pLSA)[5], and Latent Dirichlet Allocation (LDA)[6]. LSA is one of the semantic analysis algorithms which combines some latent factor of textual data by adding additional vector space features such as singular value decomposition (SVD) of document-term matrix to analyze the document-term relationships. pLSA model is proposed to overcome the disadvantage found in by LSA model, trying to decrease the degree of computation by using probabilistic approach. pLSA analyzes the document-term relationships using latent topic space, just like LSA, which projects the term tj in set T together with document di in set D to a set of k latent topics Tk. pLSA and LSA try to represent the original document space with a lower dimension space called latent topic space. In Hofmann [5], P(Tk|d) is treated as the lower dimension representation of document space, for any unseen document or query, trying to find the maximum similarity with fixed P(t|Tk). Other than LSA and pLSA, the algorithm of Latent Dirichlet Allocation (LDA) is more advantageous since LDA performs even better than previous research results in latent topic detection. In fact, LDA is a general form of pLSA, the difference between LDA and pLSA model is that LDA regards the document probabilities as a term mixture model of latent topics. Girolamin and Kaban [3] shows that the pLSA model is just a special case of LDA when Dirichlet distributions are of the same.

### III. THE PROPOSED METHOD

Here we propose our classification process based on latent topic discovery for KDDCup 2008. For the huge gap between target set and outliers, we design a feature preprocessing flow for this situation:

1. According to the data distribution in each feature, applying standard outlier detection method[7,8] to detect and normalize them.
2. Simplify the data complexity by applying feature scaling methods[9].
3. Reducing data noise by a well-known information retrieval technique called TF-IDF[10].

The first part of feature preprocessing is the outlier adjustment. Statistically, outliers are observed as an extremely biased data than the normal ones. Grubb’s test defined outliers as the member outside the largest absolute deviation from the sample mean in units of the sample standard deviation[7]. The training data of KDDCup 2008 shows a heavy-tailed condition. Without knowing the meanings of features in advance, we propose to adjust the outliers by using interquartile range:

\[
Q_1 \text{ as the 25th percentile data value of target feature } \\
Q_3 \text{ as the 75th percentile data value of target feature } \\
\text{Interquartile range } IQR = |Q_3 - Q_1|
\]

\[
\text{Adjustment upper bound } AUB = Q_3 + 3 * IQR \quad (1) \\
\text{Adjustment lower bound } ALB = Q_1 - 3 * IQR \quad (2)
\]

The second part of preprocessing is the scaling of feature values. Feature scaling not only can reduce the data complexity but is also possible to discover some characteristics of data. According to [9], there are several ways to scale features to exemplify the differences between target set and outliers: scaling by variance, scaling by domain, and scaling by min-max. The scaling by variance method simply divides each feature value by pre-calculated variance. The scaling by domain method scales each feature value to an assigned range. The scaling by min-max method assigns minimum maximal feature value as the radius R of a sphere, then scaling every value to [0, R]. In our experiment, we apply variance and domain scaling to fit to our latter steps of the classification process.

The final step of feature preprocessing is noise reduction. In our previous experience [11], it is found that the noise reduction technique, TF-IDF, is able to improve classification correctness by removing highly frequent but meaningless features. The TF-IDF method is a weighted feature filtering mechanism in information retrieval domain. This statistical approach evaluates the importance of a term(feature) in a document(candidate) by their appearance in the whole data set:

\[
\text{Term frequency } \\
tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (3)
\]

\[
\text{Inverse document frequency } \\
idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \quad (4)
\]

\[
\text{(tf-idf)}_{ij} = tf_{ij} \times idf_i \quad (5)
\]

In the formula above, when a term occurrence in a document is high, the tf value will be large; when a term appears broadly among the documents, the idf value will be small. So the tf-idf value will amplify those terms with moderate high frequency term in a single document and also appear in certain amount of documents, which means pervasively occurred term will be discriminated (or called stop terms).

After feature preprocessing, the Latent Dirichlet Allocation(LDA) process is applied. We treat every candidate record as a document and the value of a feature is treated as term frequency. When the preprocessing step is finished, the result is then fed into LDA to create latent topic model. Then re-querying process for every document begins to generate topic similarities, which will be gathered as a topic vector for a document. Fig.1 shows the relationship between documents and latent topics.
The generated topic vectors are then fed into SVM[12] to create the classification model. The whole processing flow is shown in Fig. 2.

Fig. 2 The proposed processing flow

IV. EXPERIMENT

The KDDCup 2008 does not provide answers for the test set, so we will not be able to evaluate our classifier with test set data. We separate the original training data as two parts, each part contains 826 normal patient records and 59 cancerous ones. The reason to make training and test set this way is to keep the correct target set ratio. Our training data is applied outlier adjustment first, the upper bound and lower bound for each feature is calculated with formula (1) and (2) above. Every feature value above AUB will be set to AUB and every value below ALB will be set to ALB. Next step the adjusted data is applied with variance and domain scaling. Fig. 3 shows the original data distribution and preprocessed distribution.

The next step is TF-IDF calculation. We first create a TF-IDF statistics for our whole training set, as shown in Fig. 4. Then we divide the upper bound and lower bound TF-IDF value into 10 thresholds to find the best classification result. On filtering process of TF-IDF, every value below threshold will be removed, and new filtered documents generated.

These filtered documents are then fed into LDA process to generate latent topics, and topic vector for each document is calculated. Fig. 5 shows the latent topics generated by LDA model, each L_n label represents the n-th feature.
in libSVM: linear, polynomial, radial basis, and sigmoid. Among these functions, the polynomial function is selected because no cancerous mark is predicted by other kernel functions in our process. The 10 TF-IDF thresholds are applied in preprocessing step to calculate accuracy benchmarks, as shown in table I.

### Table I

<table>
<thead>
<tr>
<th>#</th>
<th>Threshold</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>99.27%</td>
</tr>
<tr>
<td>2</td>
<td>0.00000002996785829117312</td>
<td>99.32%</td>
</tr>
<tr>
<td>3</td>
<td>0.00000008062351621362247</td>
<td>99.07%</td>
</tr>
<tr>
<td>4</td>
<td>0.00004798197420076042</td>
<td>97.73%</td>
</tr>
<tr>
<td>5</td>
<td>0.0000099183082937922554</td>
<td>96.39%</td>
</tr>
<tr>
<td>6</td>
<td>0.00001774619447307921</td>
<td>98.85%</td>
</tr>
<tr>
<td>7</td>
<td>0.00003529061359145861</td>
<td>99.18%</td>
</tr>
<tr>
<td>8</td>
<td>0.00008242479532284817</td>
<td>98.68%</td>
</tr>
<tr>
<td>9</td>
<td>0.0003151423293704674</td>
<td>99.05%</td>
</tr>
<tr>
<td>10</td>
<td>0.004568104199841167</td>
<td>99.38%</td>
</tr>
</tbody>
</table>

In order to do a better evaluation of our experiment, we calculate Receiver Operating Characteristic (ROC) values and generate ROC curve to show our results. The ROC curve consists of TPR(sensitivity) as x-axis and FPR(1-specificity) as y-axis. The AUC is defined as the area under ROC curve. The AUC result under different TF-IDF thresholds is shown in table 2 and Fig.6.

### Table II

<table>
<thead>
<tr>
<th>#</th>
<th>Threshold</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.74</td>
</tr>
<tr>
<td>2</td>
<td>0.00000002996785829117312</td>
<td>0.737</td>
</tr>
<tr>
<td>3</td>
<td>0.00000008062351621362247</td>
<td>0.809</td>
</tr>
<tr>
<td>4</td>
<td>0.000004798197420076042</td>
<td>0.8945</td>
</tr>
<tr>
<td>5</td>
<td>0.0000099183082937922554</td>
<td>0.98</td>
</tr>
<tr>
<td>6</td>
<td>0.00001774619447307921</td>
<td>0.7</td>
</tr>
<tr>
<td>7</td>
<td>0.00003529061359145861</td>
<td>0.6875</td>
</tr>
<tr>
<td>8</td>
<td>0.00008242479532284817</td>
<td>0.675</td>
</tr>
<tr>
<td>9</td>
<td>0.0003151423293704674</td>
<td>0.504</td>
</tr>
<tr>
<td>10</td>
<td>0.004568104199841167</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig.6. AUC values under different thresholds (y-axis is the AUC value and x-axis represents the thresholds)

In Fig.6, we found that the 5-th TF-IDF threshold performs best in classification process.

### V. Conclusion

In this paper, our experiment faces two major challenges of KDDCup 2008: (1) extremely distributed outliers, and (2) positive samples are far smaller than negative ones. We try to propose a suitable process flow to deal with these issues and get a best AUC result of 0.98. The future improvements of our approach may lie on the following aspects:

A. Increasing the ratio of positive data: since our approach is mainly based on statistical methods, how to appropriately increase the positive sample ratio is an important way to improve the benchmark result of our approach.

B. Using patient-wise instead of point-wise processing: the proposed method in this paper is candidate-based, that is, point-wise processing. Since the number of candidates provided by every patient may vary, using point-wise processing seems to be less better than patient-wise method.

C. The use of TF-IDF approach: the TF-IDF method will be able to filter unnecessary noise contained in data set, but it is also possible to filter out important message contained in data. According to table 2, the peak performance appeared in 5th threshold. But with higher thresholds, the benchmark falls, which means positive messages contained in training data is also filtered out. How to maintain the best information for classifier will be an important issue.

### Acknowledgment

This work was supported in part by the National Science Council of Taiwan via the grant NSC 99-2221-E-156-007.

### References


