The Measurement of Distinguishing Ability of Classification in Data Mining Model and Its Statistical Significance

Lingling Zhang^{1,2,*}, Qingxi Wang¹, Jie Wei¹, Xiao Wang¹, and Yong Shi^{2,3}

¹ Graduate University of Chinese Academy of Sciences,
Beijing (100190), China

² Research Centre on Fictitious Economy and Data Science, CAS, Beijing (100190), China

³ College of Information Science and Technology, University of Nebraska at Omaha,
Omaha, NE 68118, USA
Tel.: 86-10-82680676

{zhangll, qingxiwang, Jiewei, xiaowang, yshi}@gucas.ac.cn

Abstract. In order to test to what extent can data mining distinguish from observation points of different types, the indicators that can measure the difference between the distribution of positive and negative point scores are raised. First of all, we use the overlapping area of two types of point distributions-overlapping degree, to describe the difference, and discuss the nature of overlapping degree. Secondly, we put forward the image and quantitative indicators with the ability to distinguish different models: Lorenz curve, Gini coefficient, AR, as well as the similar ROC curve and AUC. We have proved AUC and AR are completely linear related; Finally, we construct the nonparametric statistics of AUC, however, the difference of K-S is that we cannot draw the conclusion that zero assumption is more difficult to be rejected when negative points take up a smaller proportion.

Keywords: Data Mining Model, Distinguishing Ability Measurement.

1 Introduction

As for the analysis of the results of data mining and knowledge assimilation, the usual practice is presenting the results of data mining as a visual form to users by systems, and then re-analysing and re-classifying them subjectively by users providing amendment and reference for the next data mining, so as to assure the entire data mining system of more robust and higher availability[1]. Compared with the traditional analysis of data (such as query, reporting, analysis of the on-line applications), the essential difference between data mining and them is that data mining means digging information and discovering knowledge without clear assumption[2]. In other words, on the one hand the results of data mining which is a means of artificial intelligence has the "subjective" characteristic owned by systems, but on the other hand there isn't absolutely correct pre-analysis which can be used as the basis for

* Corresponding author.

comparison with the results of data mining and evaluating the quality of results of data mining, which may be the difficulty of the evaluation of predicting ability lies[3][4].

As mentioned above, in advance there is no objective and fair data processing standards, which can be used to compare with the results of data mining, therefore, a basis for the judge is bound to be the correct answer by artificially subjective evaluation. Then whether we can use the results of artificially subjective evaluation to be compared with, the answer is usually no. The data mining system with practical applications usually has to deal with numerous data. If we deal with all of the data on artificially statistical analysis, the workload is so huge that testing is often unbearable. Besides, artificially statistical analysis is so subjective that different testers' results are likely to vary from person to person, which will impact on the objectivity and impartiality of the results. Therefore, the comparing test of the results of artificial analysis is very difficult, unless in the face of the small amount of data and having enough human resources to test. But the view that the results of data mining could not be evaluated based on this is also incorrect. If the attention to the testing phase when the degree of how the results of the model match with expectations can be shifted, it will be found that testing to what extent can data mining distinguish from observation points of different types should be the real concern. We can evaluate the predicting ability of data mining model from the perspective of overlapping degree (identification degree) of the data mining results, etc.

2 Overlapping Degree and Its Statistical Test

In the following we will study the structure of overlapping degree of classification data mining algorithms and then propose the statistical tests of overlapping degree[2].

The results of some data mining usually give the probability belonging to a particular type of an observation point so as to achieve the aim of forecast. In order to make the probability more easily to be understood and to study the distribution of observing points with different probability conveniently, we will convert the probability into a continuous numerical, the level of which can show the possibility that observation point falls into any particular category vividly. For example, during the process of the application of data mining in the bank's individual credit risk assessment, customers' default probability is converted to a credit score finally[5]. The higher the score is, the greater the default probability is. In the following, we have introduced the concept of scoring to measure the forecasting ability of models. Because the score is the linear transformation of the prediction probability, the following analysis is applicable to all the assessment of data mining models based on probability.

2.1 Nature of Overlapping Degree

The thinking of overlapping degree is: the difference between distribution of positive point scores and that of negative point scores can be portrayed by the overlapping area of the two parts. First of all, considering the simplest case: both of the point scores are normal distribution, and there is only one intersection point. In this assumption, the distribution density of the two types of points can easily come out.

Fig. 1. Diagram of Overlapping Degree

Figure 1 reveals that:

$s(s)$ is used to represent abscissa of the intersection point of the two distributions. S is a critical value for dividing population and when the score of the observation point $S > s$, the observation point will be predicted as negative. We assume the means of positive point scores and negative points scores are μ_0, μ_1 and the standard deviations are σ_0, σ_1 . If it is assumed that $\mu_0 < \mu_1$ and the cumulative distribution functions of $SIY=0$ and $SIY=1$ are F_0, F_1 , the overlapping region O can be calculated by:

$$O = F_1(s) + 1 - F_0(s) \tag{1}$$

When the standard deviations of the two distributions are equal, that is, $(\sigma_0 = \sigma_1)$, there exists only one intersection point. In normal distribution, the intersection point coordinate is $s = (\mu_0 + \mu_1) / 2$.

When the standard deviations of the two distributions are not equal, that is, $(\sigma_0 \neq \sigma_1)$, there may be one intersection point or two and the abscissa of the intersection point should meet: $f_0(s) = f_1(s)$. Due to normal distribution, the equation has the same solution as formula:

$$\begin{aligned} & s^2(\sigma_1^2 - \sigma_0^2) + 2s(\mu_1\sigma_0^2 - \mu_0\sigma_1^2) + \mu_0^2\sigma_1^2 - \mu_1^2\sigma_0^2 \\ & + 2\sigma_1^2\sigma_0^2(\log(\sigma_0) - \log(\sigma_1)) = 0 \end{aligned} \tag{2}$$

If we do not assume the scores obey any certain distributions, O can be given in the form of the general non-parameters in the following:

$$o = \int \min\{f_0(s), f_1(s)\} ds \tag{3}$$

In this situation, the existence of more points is allowed. Assuming there is only one point and the relation of score S and the default probability is positive and monotone, the overlapping area is defined as:

$$O_{pos} = \min_s \{F_1(s) + 1 - F_0(s)\} \quad (4)$$

Simultaneously, when the relation of score S and the default probability is negative and monotone, the overlapping area is defined as:

$$O_{neg} = \min_s \{F_0(s) + 1 - F_1(s)\} \quad (5)$$

Therefore, when the relation of score S and the default probability is monotone, there is:

$$O_{neg} = \min\{O_{pos}, O_{neg}\} \quad (6)$$

It can be seen easily that if the two types of points are completely separated, the overlapping area O is 0 and if the two distributions are exactly the same, O is 1, therefore, there is a measure T used to instruct predictive power:

$$T = 1 - O_{mon} = \max_s |F_0(s) - F_1(s)| \quad (7)$$

The value of indicator T is 0~1. When $T=1$, it means two types of people are completely separated, and when $T=0$, it stands for the other extreme circumstances. T in the positive and negative correlation cases is as follows:

$$\begin{aligned} T_{pos} &= 1 - O_{pos} = \max_s \{F_0(s) - F_1(s)\} \\ T_{neg} &= 1 - O_{neg} = \max_s \{F_1(s) - F_0(s)\} \end{aligned} \quad (8)$$

In the monotone cases, T can be estimated by the nonparametric estimation (such as empirical distribution function) of the cumulative distribution functions F_0 , F_1 . In the assumption of normal distribution, O (or T) can be calculated by the parameters of the normal distribution. In a more general assumption of the distribution, O (or T) can be calculated by the nonparametric estimation of probability density of the distribution of scores.

2.2 Statistical Tests of Overlapping Degree

According to the formula 3.8, we have found the indicator T is the same in the form with statistics of Kolmogorov-Smirnov test, that is, both are counting the greatest vertical distance $D(T)$ of the two probability distributions of the accumulated experience. When the sample size $N > 200$, statistics D obey normal distribution. Kolmogorov-Smirnov test is used to detect differences in location, scale, partial or other aspects of the two distributions (any difference between the two distributions), so we can use the statistics to verify whether the data mining model can distinguish observation points of different types. When we choose K-S test to deal with the overlapping degree, we calculate a statistics, that is, testing statistics D based on the sample, and then compare the shape of distribution of samples with normal distribution, so as to arrive at a value p ($0 < p < 1$), that is, the actual significance level to describe of the degree of suspicion of the idea. If the p value is less than a given significance level (such as 0.05), the original assumption is so suspicious that the data is believed not from the normal distribution, while on the contrary the data is believed from the normal distribution.

We consider such assumption that:

H_0	H_1	test statistics	reject condition
(1) $F_1(x) = F_0(x)$	$F_0(x) > F_1(x)$	$\hat{T}_{pos} = \max_s \{ \hat{F}_0(s) - \hat{F}_1(s) \}$	$\hat{T}_{pos} > \Delta_{n_1, n_0; 1-\alpha}$
(2) $F_1(x) = F_0(x)$	$F_0(x) < F_1(x)$	$\hat{T}_{neg} = \max_s \{ \hat{F}_1(s) - \hat{F}_0(s) \}$	$\hat{T}_{neg} > \Delta_{n_1, n_0; 1-\alpha}$
(3) $F_1(x) = F_0(x)$	$F_1(x) \neq F_0(x)$	$\hat{T} = \max_s \hat{F}_0(s) - \hat{F}_1(s) $	$\hat{T} > \Delta_{n_1, n_0; 1-\alpha/2}$

We use the statistics from (1) (2) to test whether the distributions of F_0, F_1 are the same. The definition of refusal conditions is as follows:

$$\Delta_{n_1, n_0; 1-\alpha} = \Delta_{q; 1-\alpha} \quad q = \left\lfloor \frac{n_o \cdot n_1}{n_o + n_1} \right\rfloor$$

n_0, n_1 stand for the number of positive and negative points and α is a given confidence level. When n_1 drops, the critical value for refusing follows up, but the value of the statistics does not have such a change. This shows that when the statistics is given, the lower the ratio of the negative points is, the more difficult zero assumption is rejected.

3 Lorenz Curve, Gini Coefficient and ROC Curve

Another widely used indicator, which is the measure of the credit score performance, is AR (accuracy ratio) based on Lorenz curve and Gini coefficient. Lorenz curve is also called the choice curve, which depicts the relation between the distribution of all the observation point scores S and that of the negative point scores S ($Y = 1$) so that we can observe the different distribution points of difference through the image. Figure 2 reveals that Lorenz curve horizontal axis OP and vertical axis OL show respectively all the cumulative percentage of observation and the corresponding cumulative percentage of "negative points" after the sort by points. And diagonal OC is the average diagonal line and the broken line OPC is uneven line [2].

In order to link with the cumulative distribution, we assume a negative score:

$$R = -S$$

Then the coordinate of Lorenz curve can be expressed as:

$$\{L_1(r), L_2(r)\} = \{P(R < r), P(R < r | Y = 1)\}, \quad r \in (-\infty, \infty) \quad (9)$$

Because $P(R < r) = 1 - F(s)$, which is equivalent to

$$L(s) = \{L_1(s), L_2(s)\} = \{1 - F(s), 1 - F_1(s)\}, \quad s \in (-\infty, \infty) \quad (10)$$

Lorenz curve can be estimated by the cumulative distribution function of experience. The best Lorenz curve corresponds to the scores that can separate "positive" points from "negative" points completely:

$$L_{opt}(s) = \{1 - F(s), g(1 - F(s))\}, \quad s \in (-\infty, \infty) \quad (11)$$

Fig. 2 Diagram of Lorenz Curve

$$g(x) = \begin{cases} \frac{x}{P(Y=1)} & 0 < x \leq P(Y=1) \\ 1 & P(Y=1) < x \leq 1 \end{cases} \quad (12)$$

The corresponding scores of the Lorenz curve coinciding with the diagonal completely are completely random, which means scores have nothing to do with the level of customers. So Lorenz curve can compare the performance of categories: good classification score is close to the optimal curve and the scores with weak distinguishing ability is close to the diagonal.

Lorenz curve can depict distinguishing ability of results of data mining intuitively and Gini coefficient can quantify the performance of credit scoring, which shows 2 times of the area of the graphics surrounded by Lorenz curves and diagonal:

$$G = 2 \int_{-\infty}^{+\infty} \{1 - F_1(s)\} d\{1 - F(s)\} - 1 = 1 - 2 \int_{-\infty}^{+\infty} F_1(s) d(s) \quad (13)$$

For the optimal Lorenz curve is the optimal Gini coefficient, the next formula is given:

$$G_{opt} = P(Y=0) = 1 - P(Y=1) \quad (14)$$

In fact accuracy rate (AR) refers to the ratio of Gini coefficient and Gini coefficient when Lorenz curve is the optimal. AR is defined as:

$$AR = \frac{G}{G_{opt}} = \frac{G}{1 - P(Y=1)} \quad (15)$$

As we can see from the analysis above, when Lorenz curve is strictly convex, and the relation of credit score S and the default probability Y is monotone and positive (the higher the score is, the greater the default probability is), AR is defined between 0 and 1. Supposing the relation of credit score S and the default probability Y is

monotone and negative, AR may be negative, which means we have to add a symbol to get a positive.

A curve that is similar to the Lorenz curve is ROC curve (receiver operating characteristic curve), the coordinate of ROC curve is defined as:

$$R(s) = \{1 - F_0(s), 1 - F_1(s)\} \quad (16)$$

Compared with the Lorenz curve, the ordinates of ROC curve obey the cumulative distribution of positive points, rather than the cumulative distribution of all the observation points. When the negative points take up a very small proportion, the shape of the two curves are very similar because F roughly equals to F_0 , making Lorenz curve and ROC are very similar in shape. As the same with Lorenz curve, the optimal ROC curve corresponds to the scores that can separate the "positive" points from the "negative" ones totally, so the vertex coordinate of optimal ROC curve is (0,0) (1,0) (1,1).

4 AUC Index and Its Nonparametric Test

4.1 AUC Index

In order to quantify the difference between F_0 and F_1 , the concept of AUC (area under curve) is introduced[2]:

$$AUC = \int_{-\infty}^{+\infty} \{1 - F_1(s)\} d\{1 - F_0(s)\} = 1 - \int_{-\infty}^{+\infty} F_1(s) dF_0(s) \quad (17)$$

When $AUC = 0$, there is no difference between F_0 and F_1 , and when $AUC = 1$, the difference between F_0 and F_1 is the maximum. It should be noted that: there is linear correlation between AUC and AR, which is proved by literature [6]:

$$AR = 2AUC - 1 \quad (18)$$

From the definition of Gini coefficient, we can get:

$$\begin{aligned} \frac{1-G}{2} &= \int_{-\infty}^{+\infty} F_1(s) d(s) = \int_{-\infty}^{+\infty} F_1(s) d\{P(Y=0)F_0(s) + P(Y=1)F_1(s)\} \\ &= P(Y=0) \int_{-\infty}^{+\infty} F_1(s) dF_0(s) + P(Y=1) \int_{-\infty}^{+\infty} F_1(s) dF_1(s) \\ &= P(Y=0)(1 - AUC) + P(Y=1) \cdot \frac{1}{2} = \frac{P(Y=0)}{2} - P(Y=0) \cdot AUC + \frac{1}{2} \end{aligned} \quad (19)$$

So

$$G = 2AUC \cdot P(Y=0) - P(Y=0). \quad (20)$$

By substitution of G in $AR = G / P(Y=0)$, the result is got.

This shows that the result must be the same whether use AR or AUC to evaluate the scores of different models to be good or bad.

4.2 Nonparametric Test of AUC

In statistical tests, when the overall distribution type is known, the statistical method which uses indexes of samples to make inference or hypothesis testing of the overall parameters is called parametric test; when the overall distribution is unknown, nonparametric test can be used. Many of the classic nonparametric tests such as Wilcoxon rank test and Mann-Whitney U test, can verify whether two distributions are from the same collectivity. [7] Two-sample Wilcoxon (or Mann-Whitney) does not have the premise of normality and the only requirement is that the sample is from the same conventionally continuous distribution when the assumption is valid. Wilcoxon test is symmetric test, testing whether the symmetric center of the difference collectivity is 0, and thus infer whether the two samples are from collectivities with the same central location[8][9].

Wilcoxon rank sum test should be applied to compare the information from two samples. The basic idea is: If the test supports the assumption, the rank sum of the two groups should not make that much difference. The basic steps are:

- (1) Establish assumptions;
 H_0 : the same overall distribution of two groups;
 H_1 : the different locations of overall distribution of two groups;
- (2) Rank for the two mixed groups;
- (3) Take rank sum of the group whose sample size is the smallest as the test statistics T ;
- (4) Assuming sample size of the smaller group as n_1 , check the critical value sheet using the difference between the two sample size, that is, $n_2 - n_1$ and T value;
- (5) Make conclusions based on P value.

When the sample size is large, normal approximation should be applied to u test. When the same rank is much, correction formula should be applied to u value.

Applying the test methods to evaluate the results of data mining, we get the method to construct u statistics:

$$U = \text{number of } \{s_{i1} > s_{j0}\} \quad (21)$$

When the positive and negative points are distinguished completely, $U = n_0 n_1$ is defined, but if they cannot be distinguished at all, in other words, there is no correlation between score S and the default label variable Y , and then the probability of event $S_{i1} > S_{j0}$ is $1/2$. So $U / n_0 n_1$ statistics can be estimated as follows:

$$\tilde{U} = P\{(S | Y = 1) > (S | Y = 0)\} = \int \{1 - F_1(s)\} = AUC \quad (22)$$

So

$$U = \left(\frac{\widehat{AR} + 1}{2} \right) \cdot n_0 \cdot n_1 \quad (23)$$

When the distribution of scores is not continuous, the relation between U and AUC also set up. However, customers with the same scores mustn't be default customers, so $S_{i1} = S_{j0}$ should also be considered by U statistics.

$$P \{(S | Y = 0) > (S | Y = 1)\} + \frac{1}{2} P \{(S | Y = 0) = (S | Y = 1)\} \quad (24)$$

In 1975 Lehmann proved that based on the assumption of $F_1(x) = F_0(x)$, U obeys gradual normal distribution when the sample size is large. We can study the following assumptions:

	H_0	H_1	Test Statistic	Reject condition
(1)	$F_1(x) = F_0(x)$	$F_0(x) > F_1(x)$	U	$U > k_{n_0, n_1; 1-\alpha}$
(2)	$F_1(x) = F_0(x)$	$F_0(x) < F_1(x)$	U	$U < n_0, n_1 - k_{n_0, n_1; 1-\alpha}$

The critical value is:

$$k_{n_0, n_1; 1-\alpha} = \frac{n_0, n_1}{2} + u_{1-\alpha} \cdot \sqrt{\frac{1}{12} \cdot n_0 \cdot n_1 \cdot (n_0 + n_1 + 1)} \quad (25)$$

Known by the formula above, both the critical value and the statistics U reduce when n_1 drops, so we can not draw the conclusion that zero assumption is more difficult to be rejected when negative points take up a small proportion.

5 Conclusions

In order to test to what extent can data mining distinguish from observation points of different types, the indicators that can measure the difference between the distribution of positive and negative point scores are raised. First of all, we use the overlapping area of two types of point distributions, that is, overlapping degree, to describe the difference, and discuss the nature of overlapping degree. We found that overlapping degree is similar with K-S statistics in the form of measurement, so we use K-S statistics to examine whether the results of data mining model distinguish between “positive” and “negative” points in the given level of confidence .At the same time we have found when the statistics is determined, the smaller the negative points take up the proportion, the more difficult the zero assumption is rejected; Secondly, we put forward the image and quantitative indicators with the ability to distinguish different models: Lorenz curve, Gini coefficient, AR, as well as the similar ROC curve and AUC. We have proved AUC and AR are completely linear related; Finally, we construct the nonparametric statistics of AUC, however, the difference of K-S is that we can not draw the conclusion that zero assumption is more difficult to be rejected when negative points take up a smaller proportion.

Acknowledgements

This research has been partially supported by a grant from National Natural Science Foundation of China (#70501030, #70621001, #90718042) and Beijing Natural Science Foundation (#9073020).

References

1. Padmanabhan, B., Tuzhilin, A.: Knowledge refinement based on the discovery of unexpected patterns in data mining. *Decision Support Systems* (1999)
2. Wei, J.: Objective Measurements of Intelligent Knowledge (Master Dissertation), Graduate University of Chinese Academy of Sciences (2008)
3. Sveiby, K.E.: *The New Organizational Wealth Managing and Measuring Knowledge-based Assets*. Berrett-Koehler-Publishers, San Francisco (1997)
4. Demsar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research* 7, 1–30 (2006)
5. Infield, N.: Capitalizing on knowledge. *Information World Review* (1997)
6. Rasero, B.C.: Statistical Aspects Of Setting Up A Credit Rating System (2003)
7. Shannon, C.E.: *The Mathematical Theory of Communication*. BSTJ (1948)
8. Fahrmeir, L., Hamerle, A., Tutz, G.: *Multivariate Statistische Verfahren*. Walter de Gruyter, Berlin (1996) (in German)
9. Michie, et al.: *Machine learning, neural and statistical classification*. Ellis Horwood, Chichester (1994)