Vergleich von Erkennungsmethoden

Simon Günter

IAM-04-001

März 2004

Vergleich von Erkennungsmethoden
Comparison of Classification Methods

Simon Günter
Forschungsgruppe Bildanalyse und Künstliche Intelligenz
Institut für Informatik und angewandte Mathematik
Universität Bern

0 Vorwort

Das Ziel des vorliegenden Berichts besteht darin, Studierenden und Mitarbeiterinnen der Fachgruppe für Künstliche Intelligenz am Institut für Informatik und angewandte Mathematik der Universitt Bern einen kurzen Leitfaden zu geben, um Klassifikationsmethoden zu testen und miteinander zu vergleichen.

CR Categories and Subject Descriptors: G.3.4 [Experimental design]
General Terms: Verification
Additional Key Words: Experimental setup, statistical significance

1 Einleitung und Notation

Gegeben seien zwei Methoden zur Generierung eines Erkennungssystems, M_1 und M_2. Im einfachsten Fall sind M_1 und M_2 einfach zwei bereits generierte Erkenner. M_1 und M_2 können z.B. auch Methoden zur Generierung von Mengen von Erkennern sein, die bei der Erkennung kombiniert werden. Die beiden Systeme, die durch die Methoden erzeugt wurden, werden auf der Testmenge T getestet. Wichtig ist dabei, dass M_1 und M_2 die Menge T nicht verwendet haben. Insbesondere wurden M_1 und M_2 nicht auf T optimiert. Die Anzahl korrekt erkannter Elemente von T ist E_1 für M_1 und E_2 für M_2. Die Erkennungsraten R_1 und R_2 werden nun wie folgt definiert:

$$R_1 = \frac{E_1}{|T|},\ R_2 = \frac{E_2}{|T|} \tag{1}$$

$M_1(t)$ und $M_2(t)$ sind die Erkennungsresultate, die mit Hilfe von M_1 und M_2 für das Element t erzeugt wurden. E_{12} ist die Anzahl Elemente von T, für die $M_1(t)$ und $M_2(t)$ korrekt sind. Die "gemeinsame" Erkennungsrate R_{12} ist nun definiert als:

$$R_{12} = \frac{E_{12}}{|T|} \tag{2}$$

Im Folgenden nehmen wir an, dass wir zeigen wollen, dass M_1 "besser" als M_2 ist. Es wird nur ein Vergleich gemacht, wenn M_1 für den vorliegenden Fall wirklich besser war, d.h. falls $R_1 > R_2$.

Für den Vergleich haben wir folgende Nullhypothese: M_1 und M_2 sind gleich gut, d.h. sie haben die gleiche durchschnittliche Erkennungsrate. Falls die Nullhypothese verworfen wird, bedeutet dies, dass M_1 eine höhere durchschnittliche Erkennungsrate hat (man beachte, dass $R_1 > R_2$ vorausgesetzt wurde).

Die Nullhypothese wird verworfen, falls diese aufgrund gemessener Kenngrössen zu unwahrscheinlich ist. Die Wahrscheinlichkeit der Nullhypothese wird im Folgenden P_n genannt. Oft wird im voraus bestimmt, bei welcher Wahrscheinlichkeit die Nullhypothese verworfen wird. Die Grenzwahrscheinlichkeit wird Signifikanzniveau α genannt. Falls $P_n < \alpha$, dann wird die Nullhypothese verworfen. Alternativ kann anstatt α und der Angabe, ob die Nullhypothese verworfen wurde, direkt P_n angegeben werden. In diesem Fall wird mehr Information weitergegeben.

2 Einfacher Vergleichstest

Die gemessene Testgrösse beim einfachen Vergleichstest ist:

$$Z = \frac{R_1 - R_2}{\sqrt{\frac{1}{|T|} \cdot (R_1 \cdot (1 - R_1) + R_2 \cdot (1 - R_2))}} \tag{3}$$

Falls die Nullhypothese zutrifft und R_1 unabhängig von R_2 ist, dann ist der Erwartungswert von Z gleich 0 und die Standardabweichung 1. Für einen hohen Wert von $|T|$ ist Z nahezu normalverteilt ($|T| > 50, R_1 \cdot |T| > 2.5, R_2 \cdot |T| > 2.5$). Damit kann mit dem ermittelten Wert von Z in der Nomalverteilungstabelle P_n abgelesen werden. Da $R_1 > R_2$ vorausgesetzt wurde, wird ein einseitiger Test gemacht. Im Folgenden gibt der Wert $N(x)$ die Wahrscheinlichkeit an, dass bei einer Normalverteilung mit Mittelwert 0 und Varianz 1 ein Wert grösser als x vorkommt. In dem vorliegenden Fall gilt $P_n = N(Z)$.

Beispiel: $R_1 = \frac{2}{3}, R_1 = \frac{1}{3}, |T| = 64 \rightarrow Z = 4$. In diesem Fall ist die Wahrscheinlichkeit der Nullhypothese $P_n = 0.0033\%$, dass heisst sogar bei einem Signifikanzniveau von $\alpha = 0.1\%$ wird die Nullhypothese verworfen.

Die Herleitung der Formeln basiert zu einem grossen Teil auf [3] und [1].

2.1 Fortgeschrittener Vergleichstest

Beim einfachen Vergleichstest wird die Unabhängigkeit der Resultate der beiden Methoden angenommen. In den meisten Fällen sind diese Resultate jedoch korreliert (d.h. R_{12} ist wesentlich höher als $R_1 \cdot R_2$.). Bei korrelierten Resultaten ist ein Unterschied der Erkennungsrate der Methoden signifikanter als bei Unabängigen.

Ein Test, der die Korrelation berücksichtig, ist der gepaarte t-Test. Für diesen Test wird eine neue Zufallsvariable X eingeführt. Der Wert von X für ein Testelement t ist durch die untenstehende Formel definiert:

$$X(t) = \begin{cases} 1 & \text{falls } M_1(t) \text{ korrekt und } M_2(t) \text{ nicht korrekt ist} \\ 0 & \text{falls } M_1(t) \text{ und } M_2(t) \text{ beide korrekt oder beide nicht korrekt sind} \\ -1 & \text{falls } M_2(t) \text{ korrekt und } M_1(t) \text{ nicht korrekt ist} \end{cases} \tag{4}$$

Es gilt zu beachten, dass immer Resultate des gleichen Testelements miteinander verglichen werden.

Falls die Nullhypothese zutrifft und $|T|$ genügend gross ist ($|T| > 30$), ist X normalverteilt. Somit kann eine neue Testgrösse eingeführt werden:

$$Z = \frac{\mu_x}{\sqrt{\frac{\sigma_x}{|T|}}} \tag{5}$$

Dabei ist μ_x der Mittelwert und σ_x die Varianz von X. Es gilt $\mu_x = R_1 - R_2$. Für die Berechnung von von σ_x kann R_{12} verwendet werden

$$\sigma_x = (R_1 - R_{12}) \cdot (1 - R_1 + R_2)^2 + (R_2 - R_{12}) \cdot (1 + R_1 - R_2)^2 + (1 + 2 \cdot R_{12} - R_1 - R_2) \cdot (R_1 - R_2)^2 \tag{6}$$

Die Testgrösse Z ist normalverteilt mit Mittelwert 0 und Varianz 1. Mit einer Nomalverteilungstabelle kann nun P_n abgelesen werden (auch hier wird ein einseitiger Test verwendet), d.h. $P_n = N(Z)$.

Beispiel: $T = 100, R_1 = 0.5, R_2 = 0.6$. Mit M_1 werden alle 50 Elemente richtig erkannt, die auch mit M_2 richtig erkennt werden $\rightarrow \mu_x = 0.1, R_{12} = 0.5$. $\sigma_x = 0.1 \cdot 0.9^2 + 0 \cdot 1.1^2 + 0.9 \cdot 0.1^2 = 0.09 \rightarrow Z = \frac{0.1}{\sqrt{\frac{0.09}{100}}} = 3.33 \rightarrow P_n = 0.06\%$. Wendet man die einfache Vergleichsmethode an, so bekommt man $Z = \frac{0.1}{\sqrt{\frac{1}{100} \cdot (0.24+0.25)}} = 1.43$ und somit $P_n = 7.78\%$.

Die Herleitung der Formeln basiert zu einem grossen Teil auf [1].

2.2 Vorzeichentest

Oft werden Erkennungsmethoden M_1 und M_2 mehrmals nacheinander mit verschiedenen Parameter getestet. Der variierte Parameter kann z.B. die Trainingsmenge sein oder ein Initialisierungsparameter der Methoden. Die Erkennungsraten von M_1 sind im Folgenden mit $R_{11}, R_{12}, \ldots, R_{1n}$ bezeichnet, diejenige von M_2 mit $R_{21}, R_{22}, \ldots, R_{2n}$. Die Anzahl Tests ist also n. Die Methoden können nun auf zwei Arten verglichen werden:

1. Die Resultate aller Anwendung einer Methode werden zu einer einziger Resultatenmenge zusammengefügt. Dies ist oft nicht wünschenwert, da die Grössen der Testmengen sehr unterschiedlich sein können und somit Experimente mit grossen Testmengen den Vergleich dominieren.

2. Anwendung des Vorzeichentests [2].

Beim Vorzeichentest [2] wird die Anzahl Z der Tests gezählt, bei denen $R_{1i} > R_{2i}$. Gilt die Nullhypothese, so sollte Z binominalverteilt sein mit $p = 0.5$. Mit Hilfe einer Tabelle oder der exakten Formel der Binominalverteilung kann nun P_n berechnet werden (auch hier wird ein einseitiger Test verwendet). Die exakte Formel für diesen Fall ist:

$$P_n = \sum_{i=Z}^{n} \left(\begin{array}{c} n \\ i \end{array} \right) \cdot \frac{1}{2}^i \tag{7}$$

Beispiel: Bei allen 4 Tests erzielte M_1 immer die besseren Resultate als M_2. $P_n = \frac{4!}{4!0!} \cdot 0.5^4 = 6.25\%$.

Es gilt zu beachten, dass eine Voraussetzung des Vorzeichentest die Unabängigkeit der Tests ist. Ein Beispiel, bei dem die Tests nicht unabhängig sind, ist das folgende: Gegeben sei eine Testmenge S mit 100 Elementen mit $a, b \in S$. Beide Methoden M_1 und M_2 werden nun auf die beiden Testmengen $T \backslash \{a\}$ und $T \backslash \{b\}$ angewandt (es gibt also vier Erkennungsraten $R_{11}, R_{12}, R_{21}, R_{22}$). Die Resultate

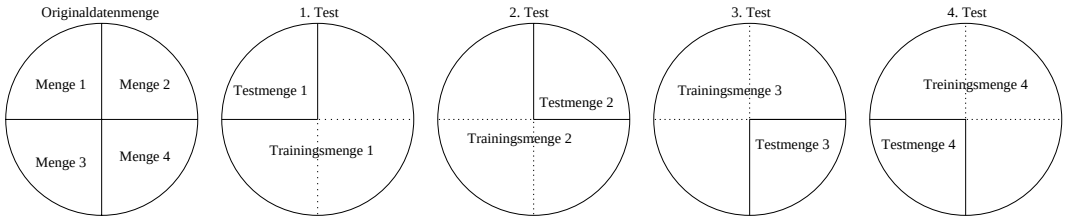

Abbildung 1: Vierfache Crossvalidierung

einer Methode werden auf beiden Testmengen fast gleich sein. In dem meisten Fällen kann jedoch die Unabängigkeit vorausgesetzt werden.

3 Crossvalidierung

Manchmal ist eine einfache Aufteilung in eine Trainings-und Testmenge schwierig, da nur wenige Daten zur Verfügung stehen. Ist die Trainingsmenge zu klein, so kann das Erkennungssystem nicht richtig trainiert werden. Ist die Testmenge zu klein, so lassen sich keine statistisch signifikante Vergleiche von Methoden durchführen. Eine Lösung dieses Problem ist die Crossvalidierung.

Bei der Crossvalidierung wird die Originaldatenmenge in n Teile aufgeteilt. Jeder dieser n Teile wird einmal zum Testen verwendet, während die anderen $n-1$ Teile für das Training gebraucht werden. Eine vierfache Crossvalidierung ist in Fig. 1 gezeigt. Die Resultate aller Tests werden am Schluss zu einer einzigen Resultatmenge zusammengefasst.

4 Zusammenfassung

Es sollen Methoden M_1 und M_2 verglichen werden. Die Erkennungsraten von M_1 und M_2 sind R_1 und R_2. Die Testmenge sei T. Der Anteil der Elemente von T für die beide Methoden korrekte Resultate liefern, ist R_{12}. P_n ist die Wahrscheinlichkeit der Nullhypothese, dass M_1 gleich gut ist wie M_2. $N(x)$ ist die Wahrscheinlichkeit, dass bei einer Normalverteilung mit Mittelwert 0 und Varianz 1 ein Wert grösser als x vorkommt.

Einfacher Test: $P_n = N(Z)$ mit

$$Z = \frac{R_1 - R_2}{\sqrt{\frac{1}{|T|} \cdot (R_1 \cdot (1 - R_1) + R_2 \cdot (1 - R_2))}} \tag{8}$$

Fortgeschrittener Test: $P_n = N(Z)$ mit

$$Z = \frac{R_1 - R_2}{\sqrt{\frac{\sigma_x}{|T|}}} \tag{9}$$

$$\sigma_x = (R_1 - R_{12}) \cdot (1 - R_1 + R_2)^2 + (R_2 - R_{12}) \cdot (1 + R_1 - R_2)^2 + (1 + 2 \cdot R_{12} - R_1 - R_2) \cdot (R_1 - R_2)^2 \tag{10}$$

Vorzeichentest: R_{ij} ist j-te Erkennungsrate von M_i, Z ist Anzahl der Fälle mit $R_{1j} > R_{2j}$ und n die Anzahl Tests. Es gilt:

$$P_n = \sum_{i=Z}^{n} \left(\begin{array}{c} n \\ i \end{array} \right) \cdot \frac{1}{2}^i \tag{11}$$

Literatur

[1] *Formeln und Tafeln, Mathematik - Physik.* Orell Füssli Verlag, Zürich, 2001. ISBN: 3280021626.

[2] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Research and Development in Information Retrieval*, pages 329–338. 1993.

[3] N. Parihar and J. Picone. Aurora working group: DSR front end LVCSR evaluation AU/384/02. Technical report, Institute for Signal and Information Processing, Department of Electrical and Computer Engineering,Mississippi State University, Mississippi, USA, 2002.