

# Kmeans pro Poker Hands

Martin Lukeš

TABLE I  
TAULKA ATRIBUTŮ

Atribut	Význam	Rozsah
C1	Barva první karty	1-4
S1	Hodnota první karty	1-13
C2	Barva druhé karty	1-4
S2	Hodnota druhé karty	1-13
C3	Barva třetí karty	1-4
S3	Hodnota třetí karty	1-13
C4	Barva čtvrté karty	1-4
S4	Hodnota čtvrté karty	1-13
C5	Barva páté karty	1-4
S5	Hodnota páté karty	1-13
Hand Power	Síla ruky	0-11

**Abstrakt— Rozpoznání síly ruky v pokeru spolu s pozorováním tellů (manýr, výraz, emoce či zvyk vyjadřující informaci o stavu ruky hráče) oddělují špatné hráče od dobrých. Tento dokument se zaměřil na zkoumání síly jednotlivých rukou.**

## I. ASSIGNMENT

Vyberte si data, která jsou použitelná pro klasifikaci nebo regresi. Data předzpracujte a nainportujte do DM aplikace. Zvolte si shlukovací algoritmus (k-means, hierarchické shlukování, SOM). Najděte nastavení parametrů zvoleného algoritmu tak, aby produkoval co nejlepší výsledky. Interpretujte výsledky.

## II. INTRODUCTION

Jako shlukovací algoritmus jsem si vybral K-means. Data byla již v číselné podobě a bylo nutné je jen normalizovat. Atributy a jejich význam je zachycen v tabulce 1. Poslední atribut Hand Power je tzv. učitel, který vyjadřuje sílu dané ruky. Položek v datech je přes jeden milion, proto jsem použil náhodný vzorek asi 3000 instancí, který pro mé účely postačil.

## III. METHODOLOGY

K-means je shlukovací algoritmus, který je zařazován do skupiny tzv. centroidních algoritmů. K-means nahrazuje všechny shluky jejich reprezentanty (centroidy) a nemusí tak vypočítávat všechny vzdálenosti, ale jen vzdálenosti všech bodů od centroidů.

K měření vzdáleností můžeme využít různých měřících metod (metrik). V pokusech byla použita eukleidovská, manhattanská a cosinova metrika. Eukleidovská -  $E(P,Q)$  - metrika měří vzdálenost dvou bodů ( $P$  a  $Q$ ) v  $n$ -rozměrném prostoru jako odmocninu z součtu kvadrátů rozdílů odpovídajících souřadnic.

TABLE II

TABULKA POROVNÁVAJÍCÍ VZDÁLENOSTI PODLE HODNOTY SILHOUETTE

Počet shluků	Eukleidovská vzd	Manhattanská vzd	Cosinova vzd
2	0,17	0,09	0,17
5	0,15	0,07	0,16
10	0,15	0,06	0,15
15	0,15	0,07	0,16
20	0,15	0,07	0,16

$$P = (p_1, p_2, \dots, p_n)$$

$$Q = (q_1, q_2, \dots, q_n)$$

$$E(P, Q) = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

Manhattanská metrika -  $M(P, Q)$  - měří vzdálenost dvou bodů ( $P$  a  $Q$ ) v  $n$ -rozměrném prostoru jako součet absolutních hodnot rozdílů odpovídajících souřadnic.

$$P = (p_1, p_2, \dots, p_n)$$

$$Q = (q_1, q_2, \dots, q_n)$$

$$M(P, Q) = \sum_{k=1}^n |p_k - q_k|$$

Kosinova metrika -  $\text{dist}(P, Q)$  - je hodnota podobnosti (úhlu), který svírají  $P$  a  $Q$ .

$$P = (p_1, p_2, \dots, p_n)$$

$$Q = (q_1, q_2, \dots, q_n)$$

$$\begin{aligned} \text{dist}(P, Q) &= \arccos \frac{\sum_{k=1}^n p_k * q_k}{\sqrt{\sum_{k=1}^n p_k^2 * \sum_{k=1}^n q_k^2}} = \\ &= \arccos \left( \frac{P * Q}{|P| * |Q|} \right) \end{aligned}$$

Na Figure 1 vidíme porovnání všech tří metrik. Nezávislá proměnná je počet centroidů a závislá je hodnota silhouette. Data jsou nejlépe shlukovaná pro dva shluky u všech metrik. Na Figure 2, 3, 4 vidíme grafy pro nejlepší počty shluků (2).

## IV. CONCLUSION

Při měření jsem zjistil, že model je nejlepší pro počet shluků roven 2 u všech metod, což je trochu zvláštní. Je možné, že data nejsou pro K-means úplně ideální, protože jeden parametr je nejlepší pro všechny modely, navíc pro dva s absolutně stejnou hodnotou silhouette. Nejlépe tedy dopadl model Cosinovy a Eukleidovské vzdálenosti.

## REFERENCES

- [1] Materiály z přednášek.

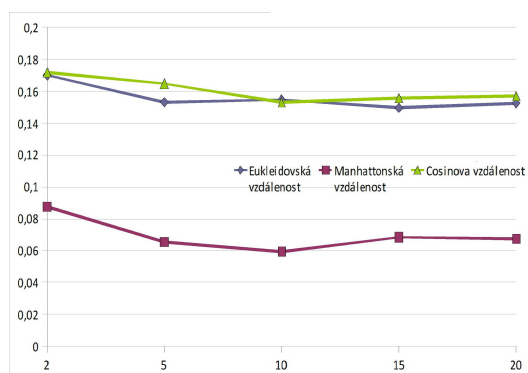


Fig. 1. Graf porovnání metrik podle hodnoty silhouette

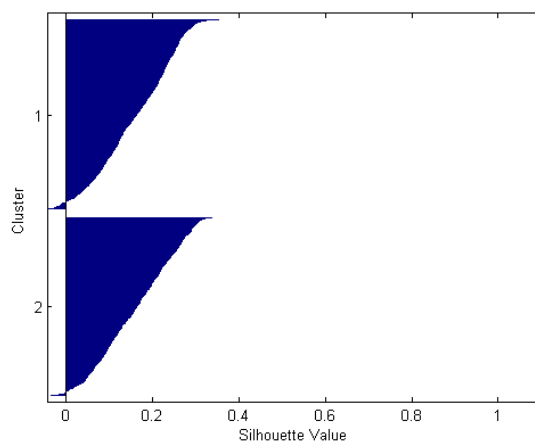


Fig. 3. Silueta Cosinovy vzd

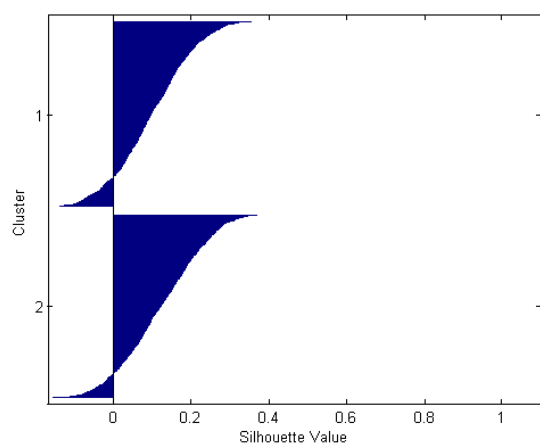


Fig. 2. Silueta Manhattanové vzd

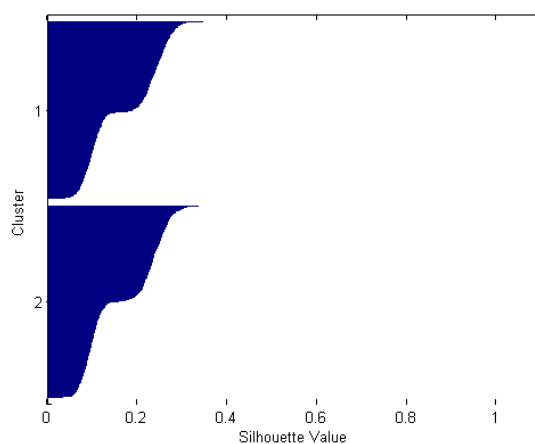


Fig. 4. Silueta Eukleidovské vzd