Datamining - Encyclopedia Academia

Datamining	
Richting	Informatica
Jaar	MINF

Bespreking

Dit vak gaat over het proces van automatisch informatie zoeken in grote hoeveelheden data. Het is een vak dat je zeker interessant zal vinden als je de intentie hebt om de master “Databases” te volgen. Er worden drie onderdelen uitvoerig besproken: “Frequent Pattern Mining”, “Classificatie” en “Clustering”. Het eerste onderdeel toont manieren hoe je een frequente itemset (verzameling items) kan vinden in een verzameling van itemsets. Een veel gebruikt voorbeeld in dit onderdeel is het winkelkarretje. Het winkelkarretje stelt de producten voor die een klant in de winkel koopt. Het winkelkarretje is dus een deelverzameling van de producten die in de winkel te koop zijn. Via Frequent Pattern Mining zal je dan proberen de producten te vinden die het meest samen met een ander product gekocht worden (en dus in een winkelkarretje voorkomen). Zo kan je misschien relaties bekomen van de vorm: bier komt vaak voor met pindanootjes. Deze informatie kan interessant zijn voor de winkel om de verkoop te stimuleren.

Classificatie en clustering bieden methoden om, gegeven een verzameling data, te beslissen tot welke klasse een item behoort. Het verschil tussen classificatie en clustering is dat bij clustering de labels van de klassen en het aantal klassen zelf gevonden worden. Bij classificatie zullen deze op voorhand bekend moeten zijn.

Puntenverdeling

Onbekend.

Examenvragen

Academiejaar 2021 - 2022 - 1ste zittijd

Bestand:ExamDataMining2021-2022 1e.pdf

Academiejaar 2020 - 2021 - 1ste zittijd

Dit was een openboekexamen, zowel het boek als notities, oefeningen en uitgeprinte dia's waren toegestaan.

1. **Pattern Mining (3 points)**
    1. Explain how the concept of Closed Itemsets helps in solving the frequent itemset mining problem.
    2. Given all closed frequent itemsets and their supports, can I still find out the support of all the not-closed frequent itemsets without going back to the data.
2. **Recommender Systems (3 points):** Explain the problem of bias in recommender systems, and how it is being handled in Matrix factorization.
3. **Classification (6 points)** Evaluate the following classification algorithms with respect to their ability to:
    - Deal with random noise;
    - Avoid overfitting;
    - Deal with meaningless attributes; that is: attributes that are independent from the class attribute and hence are useless for the classifier.

    The algorithms:

    - kNN for different values of k (that is: for higher k, is the algorithm better suited to deal with the issue mentioned above, or less);
    - Naïve Bayes;
    - Decision tree induction algorithms like Hunt's algorithm;
    - Logistic regression.

    Give an explicit answer *for each of the 12 combinations* (3 characteristics for each of the 4 algorithms) *Give a short rationale (3-4 sentences max); just answering yes/no is insufficient.*
4. **Measering Performance (2 points)** Which of the following is true? *Prove or give a counterexample.*
    1. A classifier with an accuracy of 100% always has an AUC of 1.
    2. For any two classifiers C1C1 and C2C2 it holds that: *If C1C1 has a higher accuracy that C2C2, on a test set DtDt then the AUC of C1C1 is higher than the AUC of C2C2 on DtDt as well.*

5. **Class independence (3 points)** The derivation of the Naïve Bayes classifier highly depends on the assumption that the attributes in the data are class independent.
    1. What does class independence mean?
    2. Give an *example* that illustrates the difference between independence of attributes and class independence.
    3. What goes wrong if the class independence condition is not fulfilled in yout data but you are using the Naïve Bayes algorithm nevertheless?
6. **Clustering (3 points)** Consider the k-means clustering algorithm. Recall that this algorithm starts with random centroids and concists of the repetition of two consecutive steps:
    - Assigning points to the nearest cluster center
    - Find new cluster centers

    The step of selecting the new cluster centers depends on the distance function used; for instance, if the distance function is the Euclidian distance, the mean of the points is selected, if the distance function is the Manhatten distance, the median of the points is selected.

    1. *Explain* why it is important to the use the correct combination of a distance function and method to find the cluster center.
    2. Suppose that instead of the regular Euclidian distance, we use a weighted Euclidan distance; that is, for given weights w1, ..., wd, we compute the distance between points **x**=(x1,...,xd) and **y**(y1,...yd) as follows
    ∑j=1dwj(xj−yj)2−−−−−−−−−−−−− √
    ∑j=1dwj(xj−yj)2

    Which steps of the k-means algorithm do you have to adapt to deal with this distance function and how? *Explain your answer; show derivations if applicable.*

Academiejaar 2019 - 2020 - 1ste zittijd

1. PM (SPT)

    17 mining frequent itemsets in a given dataset using apriori never seems to end:

    1. What could be the reason? (give 2 possible reasons)
    2. How could these be solved? (at least 1 for each reason)
    3. What are the main disadvantages for these solutions (if any)?

2. Classification (SPT)

    Dataset:

Age	BMI	System BP	???	Hospitalized	P(Hosp)
91	20	135	90	yes	0,73
25	27	125	75	no	0,08
87	33	150	90	yes	0,98
14	18	110	60	no	0,20
18	35	140	85	no	0,35
34	35	153	80	yes	0,69
89	??	110	65	yes	0,27
85	27	120	70	no	0,78
43	24	110	60	yes	0,06
55	29	135	75	no	0,56

    Logistic regression model:

    P(Hosp= yes|Age,BMI, SBP, DBP) = ???

Academiejaar 2008 - 2009 - 1ste zittijd

1. Short Questions

    1. What is overfitting? Briefly describe one method to prevent overfitting.
    2. What is the missing data problem? Briefly describe three methods to handle missing data.
    3. State the key difference between classification and clustering.
    4. Explain what is meant by “the curse of dimensionality”.

2. Frequent Itemset Mining

    Given the following transaction database over {a,b,c,d,e}{a,b,c,d,e}. Assume that the minimum support threshold is 40% and that the order over the items used is the lexicographical order, i.e. a<b<c<d<ea<b<c<d<e.

TID	Items bought
1	{a,b,c}{a,b,c}
2	{a,b,c,d}{a,b,c,d}
3	{b,c,d,e}{b,c,d,e}
4	{a,c,d,e}{a,c,d,e}
5	{d,e}{d,e}

    1. Derive all candidates sets in the order generated by the Apriori algorithm.
    2. Derive all candidates sets in the order generated by the Eclat algorithm.
    3. Explain the difference between the number of candidates generated by Apriori and Eclat.
    4. Show that every maximal frequent itemset is also closed.

3. Classification

    Consider the following confusion matrix, produced by a classifier TT:

    > {4}{|c|}{Predicted Class}
    >
    > & & 0 & 1
    > True Class & 0 & 910 & 70
    > & 1 & 20 & 0

    1. Compute the accuracy score obtained by the classifier TT.
    2. Would you recommend the usage of TT for this dataset? Motivate your answer.
    3. Briefly describe at least one technique that can be used to handle the “class imbalance problem”.
    4. Explain the key assumption underlying the naïve Bayes classifier.

4. Clustering

    Consider the following points in a plane:
    A1(2,10),A2(2,5),A3(8,4),B1(5,8),B2(7,5),B3(6,4),C1(1,2),C2(4,9)A1(2,10),A2(2,5),A3(8,4),B1(5,8),B2(7,5),B3(6,4),C1(1,2),C2(4,9)

    1. Compute the k-means algorithm for k=3k=3 and the Euclidean distance. Suppose that A1,B1,C1A1,B1,C1 are selected as the initial centroids.
    2. Suppose a dataminer is not satisfied with the previous obtained result. In order to obtain better clusters, he varies the kk parameter and selects the optimal clustering accordin to the optimization criterion used. What do you think of this approach?
    3. Suggest an other manner in which the clustering result can be improved by using the k-means algorithm.

Academiejaar 2007 - 2008 - 1ste zittijd

Theorie

1.
    1. Prove that maximal frequent itemsets are closed
    2. Draw an itemset lattice representing the data given in the folowing table. Assume that the support threshold is equal to 30%. Label each node in the lattice with the folowing letter(s):
        1. if the node is a maximal frequent itemset
        2. if the node is a closed frequent itemset
        3. if the itemset is frequent but neither maximal nor closed
        4. if the itemset is infrequent

TID	Items bought
1	{a,b,d,e}{a,b,d,e}
2	{b,c,d}{b,c,d}
3	{a,b,d,e}{a,b,d,e}
4	{a,c,d,e}{a,c,d,e}
5	{b,c,d,e}{b,c,d,e}
6	{b,d,e}{b,d,e}
7	{c,d}{c,d}
8	{a,b,c}{a,b,c}
9	{a,d,e}{a,d,e}
10	{b,d}{b,d}

2. Explain
    1. the nearest neighbor classification algorithm
    2. how its sensitivity to outliers is being solved
3.
    1. Explain how the monotonicity of the support (frequency) of itemsets is used in frequent set mining algorithms (such as e.g. Apriori)
    2. In tile mining, the area of a tile is not (anti-)monotone. Which technique is being used such that we can still use a search space traversal similar to frequent itemset miners.
4. Bewijs van een variatie op het apriory model waarbij partitioning wordt toegepast. (dus dat een frequente itemset in de database ook frequent moet zijn in minstens 1 van de partities) + geef minstens 1 voor- en nadeel van deze techniek.
5. Hiërarchical clustering: bereken de agglomeratieve aanpak + bespreek minstens 2 manieren om de similarity tussen 2 clusters te bepalen.

Praktijk

1. Cluster the following eight points (with (x;yx;y) representing their position in the plane) into three clusters with the k-means algorithm; using Euclidean distance as the similarity measure.

    > A1(2;10);A2(2;5);A3(8;4);B1(5;8);B2(7;5);B3(6;4);C1(1;2);C2(4;9)A1(2;10);A2(2;5);A3(8;4);B1(5;8);B2(7;5);B3(6;4);C1(1;2);C2(4;9)

    Suppose we initially assign A1A1, B1B1 and C1C1 as the cluster centers. Give the three cluster centers after every round of execution of the algorithm.

2. Een oefening, waarbij er twee klasses zijn + en - , er is slechts 1 attribuut (een reëel getal) en dan moet je de 1-, 3-, 5- en 9-neighbours bepalen van een gegeven getal op 2 manieren : majority vote mechanism en weighted-distance voting mechanism.
3. Gegeven een kleine database à la market bucket, geef de frequente sets en toon aan hoe eclat aan die sets komt.