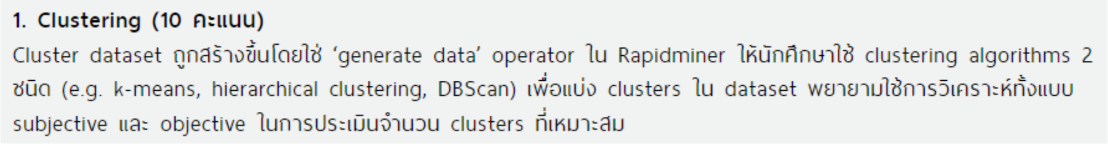
Advanced Data Mining Algorithms Practical Report

By Natibundit Suntraranusorn



1. **Clustering (10 คะแนน)**

Cluster dataset ถูกสร้างขึ้นโดยใช้ 'generate data' operator ใน Rapidminer ให้นักศึกษาใช้ clustering algorithms 2 ชนิด (e.g. k-means, hierarchical clustering, DBScan) เพื่อแบ่ง clusters ใน dataset พยายามใช้การวิเคราะห์ทั้งแบบ subjective และ objective ในการประเมินจำนวน clusters ที่เหมาะสม

Part 1: Clustering

Introduction to the dataset and its meta data.

The given data was created from the generate data operator, which is used for generating an Example Set based on numerical attributes; whereby, it is able to customize the number of attributes, number of examples, lower and upper bounds of attributes, and target function can be specified by the user. This sample Cluster dataset has 4 numerical attributes that consist of "Att1 to 4". and 500 examples were contained. Furthermore, the dataset was explored the Presenting distribution of data length by histogram in each attribute that show in the figure1. The Att2 distribution of data length is obviously similar aspects with the Att4 attribute. On the other hand, "Att1" and "Att3" attributes are individual distribution of data length.

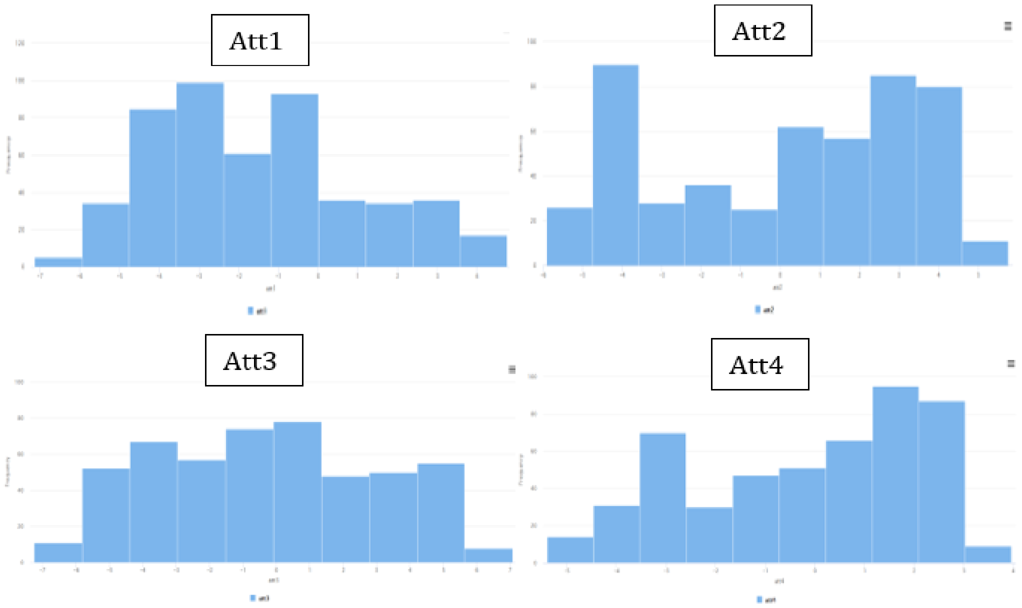


Figure 1 Distribution of data length (Histogram)

Clustering algorithms used

For clustering process, K-Means algorithm in figure2 was used for data clusters by using K = 4 due to rely on original numbers of attributes. Next steps, Performance (Cluster Distance Performance) was brought to calculate Davies **Bouldin's index** in order to search for optimizing K numbers. And then, all attributes were applied into Single Value Decomposition (SVD) in order for better visualization of the clusters.

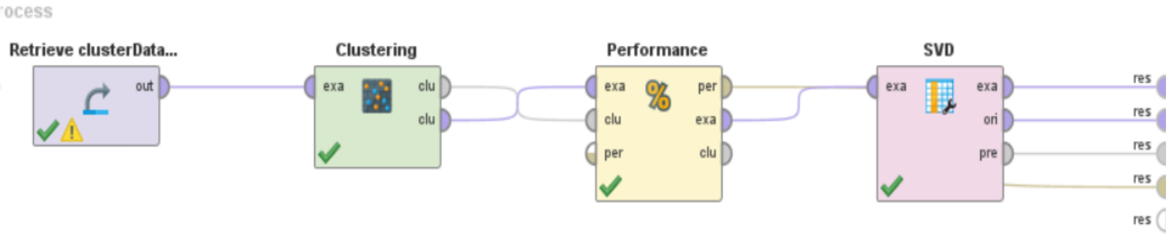


Figure 2 RapidMiner Processes (K-Mean)

Another clustering methods, DBSCAN operator in figure4 was used to separate clusters by using ESP = 5 and min point = 10. These numbers were chosen by selective ESP and min point from the scatter plot of K-distances in figure3. In this graph, If K or min points = 10 and ESP = 5, it will identify one point as either noise or border points. Data to Similarity is another operator was created to measure the different distance between two data points.

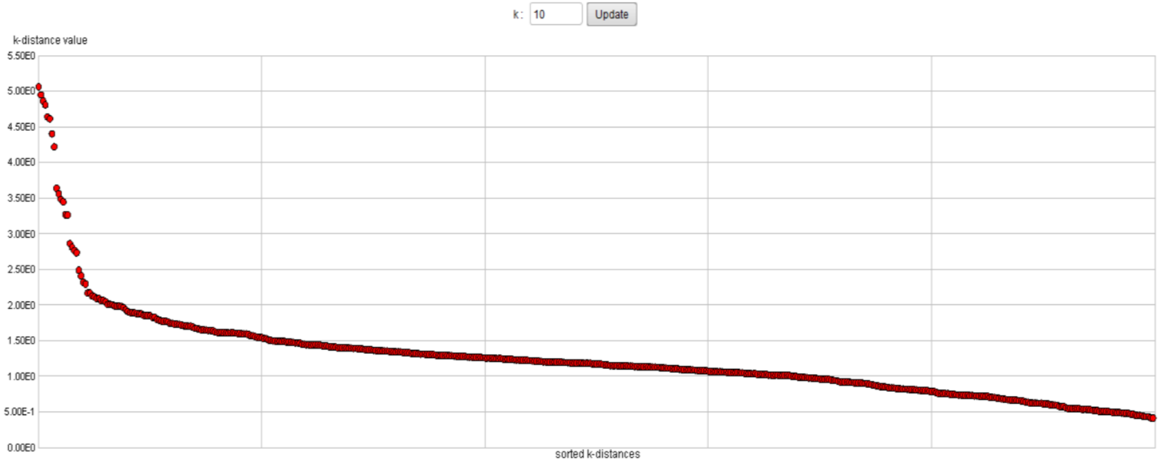

k: 10

Figure 3 Scatter plot of K-distances (Eps = 5 and Min Points = 10)

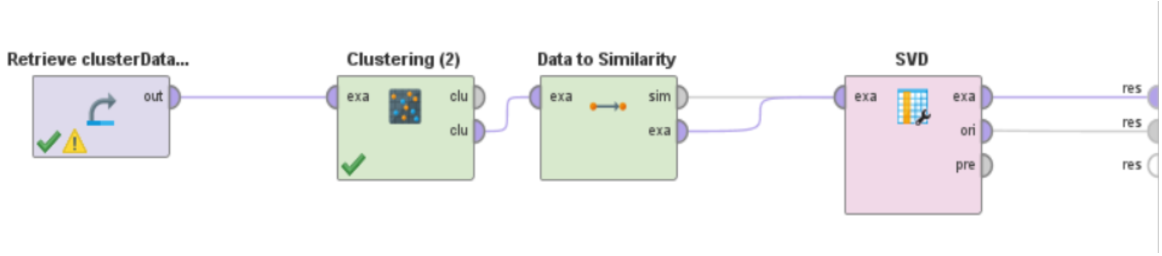


Figure 4 RapidMiner Processes (DBSCAN)

Objective measures of clusters found

Davies Bouldin's index was considered to search for optimizing K number. A smaller number of Davies Bouldin's index refers to intra-distance is more compact cluster and inter-distance is far away from each other, it is a best practice to separate clusters.

In comparison with the Davies Bouldin's number in this table below, the less Davies Bouldin's number than -0.247 or K = 4 is K = 2 and 3. Therefore, the result of K-Means with K = 2,3 and 4 will be simulated to find the cluster configuration.

K	Davies Bouldin's index
2	-0.380
3	-0.276
4	-0.247
5	-0.233
6	-0.227

The result was demonstrated in the figure 5 that separate into 4 clusters which clearly segregate with one another apart from an area of cluster_2 and 3 (Red highlight) has some black and orange data points blending together. As the same result of the figure 6 and 7, there are combining areas between two clusters.

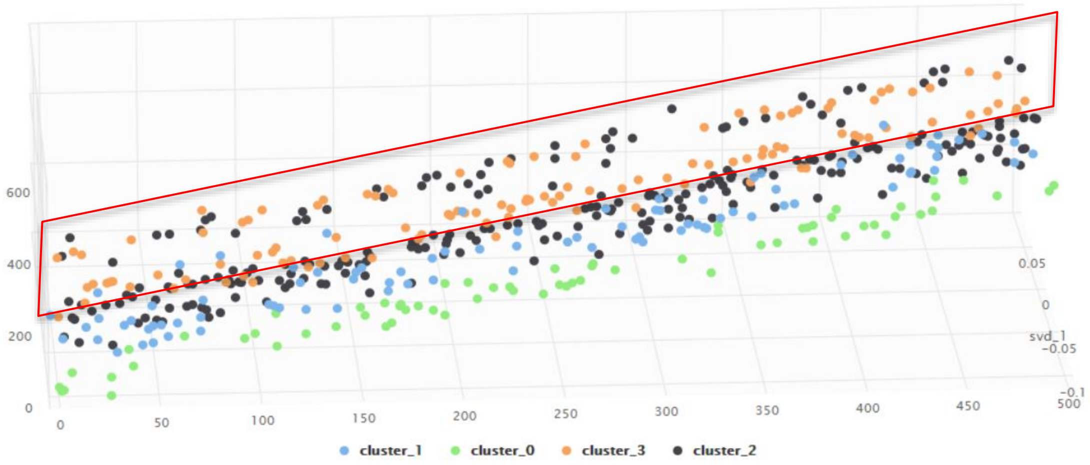


Figure 5 Clusters using K-Means with k = 4

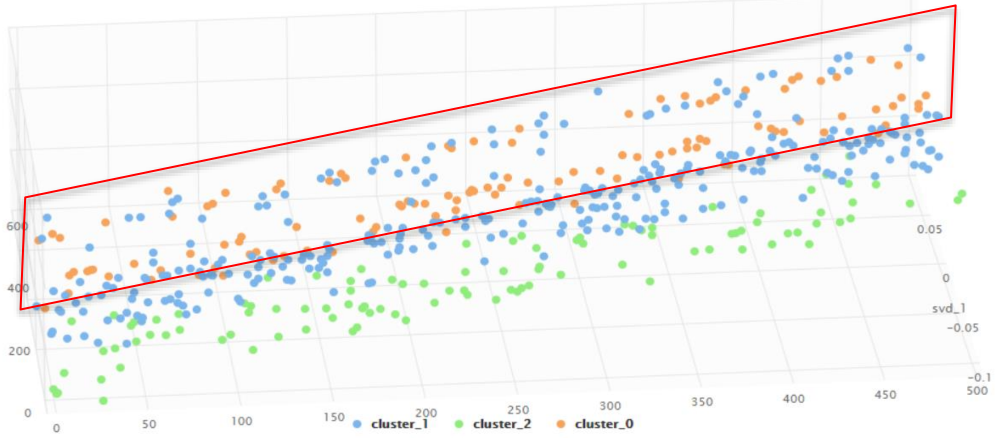


Figure 6 Clusters using K-Means with k = 3

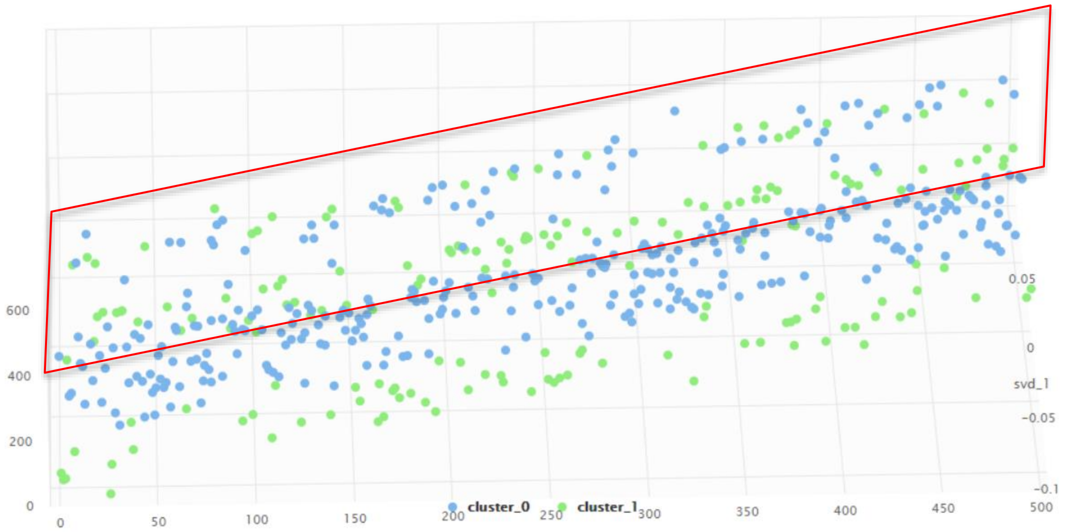


Figure 7 Clusters using K-Means with k = 2

However, the consequence of DBSCAN illustrate 2 clusters, which have Clusters_1 = 433 data points and Cluster_2 = 67 data points, it is clearly separated into two clusters.

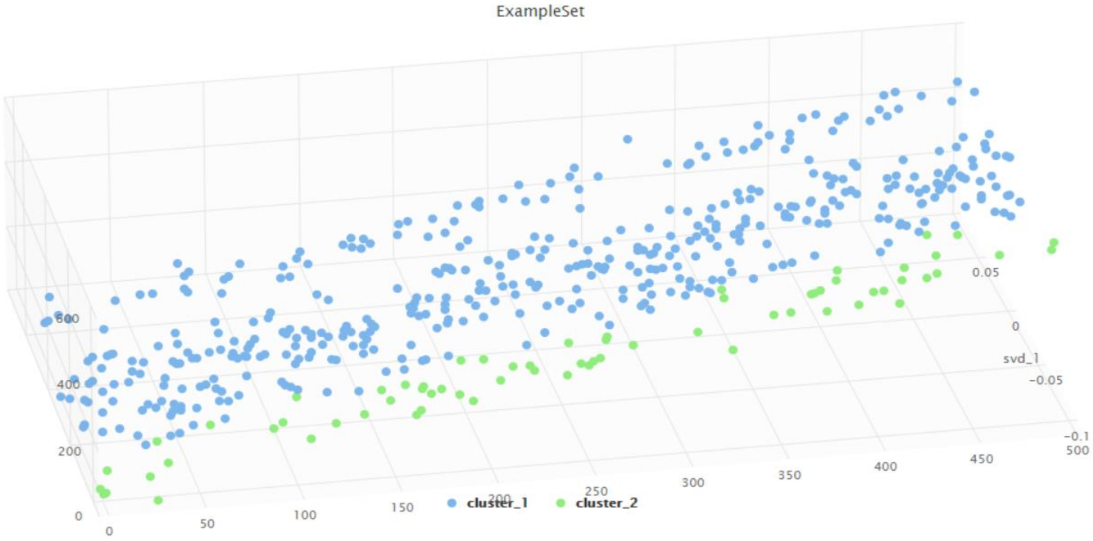


Figure 8 Clusters using DBSCAN (Eps = 5 and Min Points = 10)

Conclusion

Overall, the appropriate number of clusters is K = 2 by using DBSCAN model to arrange clusters and comparing Davies Bouldin's index to search for optimal K number despite original data have 4 attributes. As one can see in the figure 8 that is clearly visible to separate data points into two clusters greater than K-Mean method for this example dataset.

2. ANN ((10 คะแนน)

ใน adult_numeric datset วัตถุประสงค์ในการวิเคราะห์คือการทำนายว่าใครที่มีเงินเดือนมากกว่า 50K (>50k)

2.1 ให้นักศึกษาอธิบายภาพรวมของ dataset แล้วทำการ Train ด้วย ANN model ด้วย 1 hidden layer จำนวน 2 node เลือกแล้วทำการประเมินผลโดยใช้ค่า default parameter setting

2.2 ทำการปรับค่า parameter settings (learning rate, momentum, จำนวน hidden layer etc) เพื่อเพิ่ม model accuracy โดยนักศึกษาต้องเขียนอธิบายถึง parameter ที่ปรับ ว่าทำไมจึงเลือกที่จะปรับ parameter ดังกล่าว และทำการวิเคราะห์และวิจารณ์ model performance

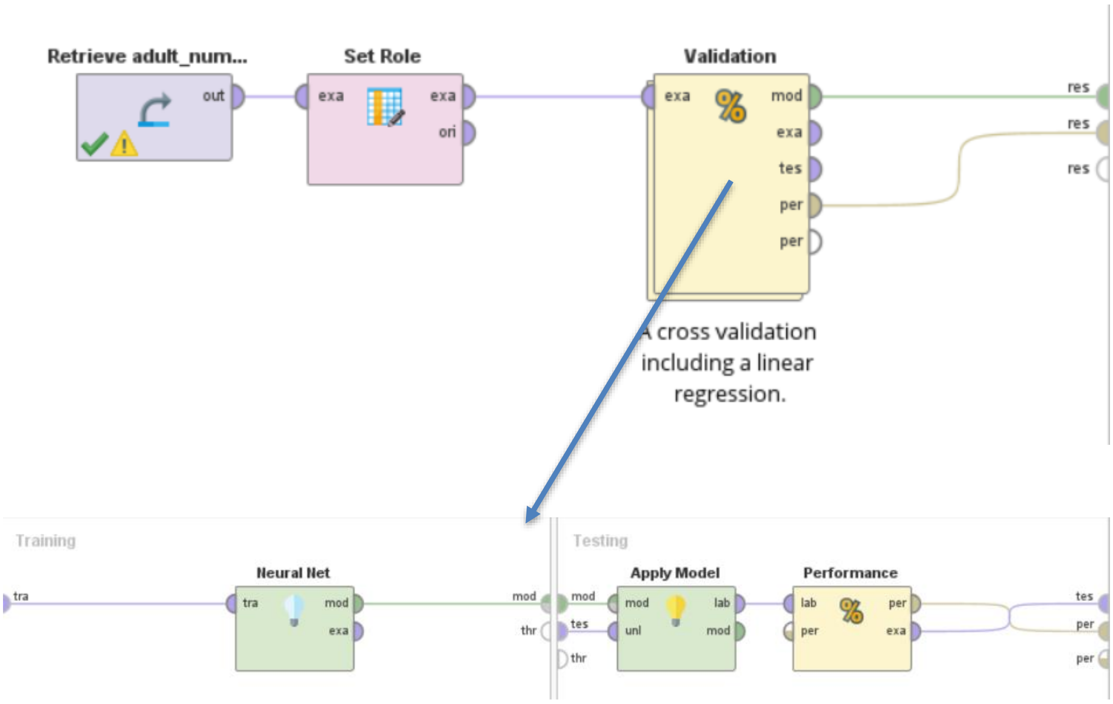
Part 2: ANN

Introduction to the dataset and its meta data.

The given dataset was contained 1,000 examples, which have 7 attributes that consist of “Age” is age of employees, “Final weight” (fnlwgt) is the weighted demography, “Education number” is the level of education, “Capital gain” is a profit from investment, “Capital loss” is a loss from investment, “Hours per week” is the time that was spent by employees per week and “Label” is an income that was separated into two categories, which has 501 numbers of labeled “<=50k”, has 499 numbers of labeled “>50k”.

Results of training using default parameters.

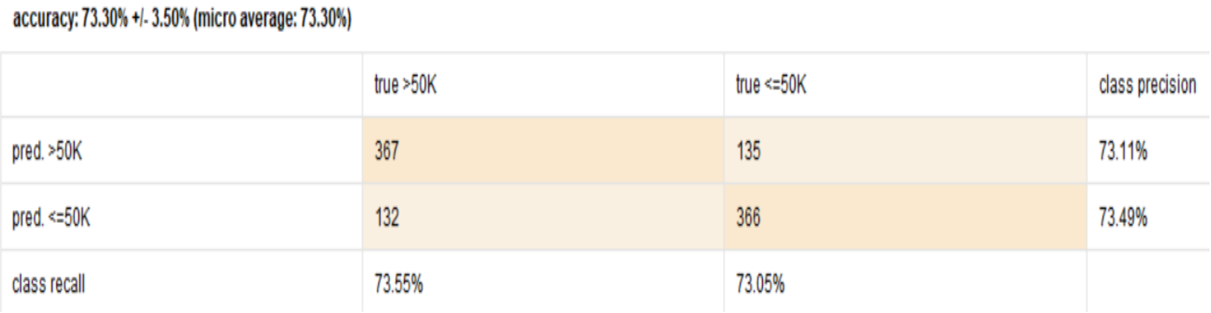
Using Set Rule Operator to identify labeled attribute from the numeric adult dataset. The Cross Validation operators (number of folds = 10) were embedded the Neural Net operator to be the main predictive model. And, the Apply model and Performance operator were used to evaluate the accuracy in figure1.



A cross validation including a linear regression.

Figure 1 RapidMiner Processes

The consequence of the default model in the figure2 has accuracy at 73.30%. As one can see, there are incorrect predictions at the similar rate occurring approximately 27% between income <=50K and income >50K. For example, the number of predictive >50K was correct at 367 all of 502, which mean the corrected prediction was 73.11% that is a True positive. On the other hands, the corrected prediction <=50K was at 366 all of 498 or 73.49% that is a True negative.



accuracy: 73.30% +/- 3.50% (micro average: 73.30%)

	true >50K	true <=50K	class precision
pred. >50K	367	135	73.11%
pred. <=50K	132	366	73.49%
class recall	73.55%	73.05%	

Figure 2 The result of Neural Net model (Default)

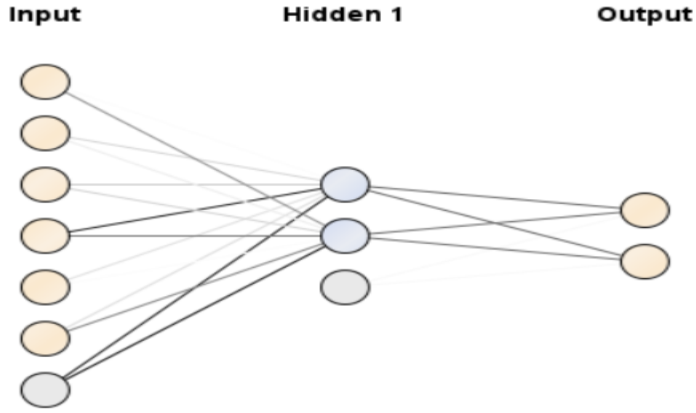


Figure 3 ANN Layers (two nodes of hidden layer)

Results when training using modified parameter settings.

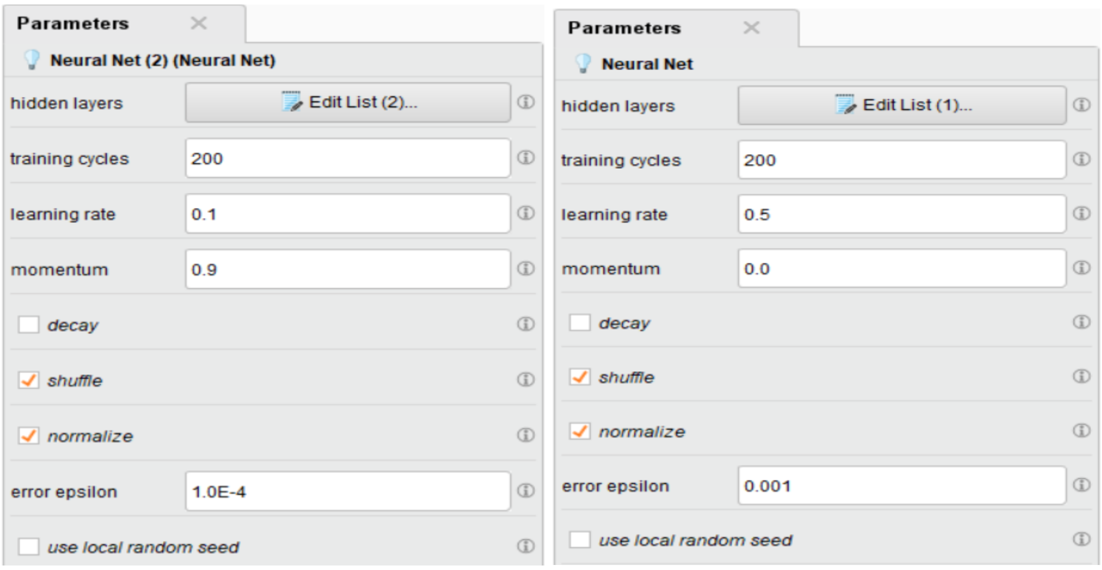
The assumption of model adjustment, momentum is an alternative parameter, which is used for adjusting in previous epochs influence weight updates to sequencing epochs. Therefore, the momentum parameter is superseded to be "zero" to avoid the complexity of setting parameters. For error epsilon, the minimum squared error is set to be 0.001, if the error rate can be reduced to 0.001, the network will be stopped. The hidden layers were tested into 3 scenarios, which have several input numbers of Training cycle and Learning rate, show the result of accuracy in the figure4.

Hidden layers	Training cycle	Learning rate	Accuracy
1.Two hidden layers with 4 nodes for first hidden layer and 2 nodes for second hidden layer	200	0.1	73.8%
		0.5	71.2%
		0.8	72.2%
	1,000	0.1	72.8%
		0.5	69.8%
		0.8	71.9%
	2,000	0.1	72.9%
		0.5	70.3%
		0.8	71.6%
2.One hidden layer with 3 nodes	200	0.1	73.3%
		0.5	73.2%
		0.8	73.6%
	1,000	0.1	72.9%
		0.5	72.9%
		0.8	73.5%
	2,000	0.1	73.8%
		0.5	72.9%
		0.8	73.0%

3.One hidden layer with 1 node	200	0.1	75.0%
		0.5	**75.7%**
		0.8	75.1%
	1,000	0.1	75.3%
		0.5	75.3%
		0.8	75.1%
	2,000	0.1	75.6%
		0.5	75.1%
		0.8	74.9%

Figure 4 Adjusting Parameters for 3 scenarios

After modified Neural Network parameters in the figure5, the result of changing hidden layers and nodes is the most essential factor that affects to accuracy in comparison with Training cycle and Learning rate parameters. The number of training cycles, which is repeated 200, 1,000 and 2,000 numbers of time, adjusts the weight of each node in order to reduce the value of the error function slightly impact to the predictive rate. Learning rate is a change of learning scale at each step that is insignificant impact in the same way of Training cycle. The consequence of adjusting model in the figure6 is slight improvement of the accuracy from 73.3% to 75.7%.



Parameters

Neural Net (2) (Neural Net)

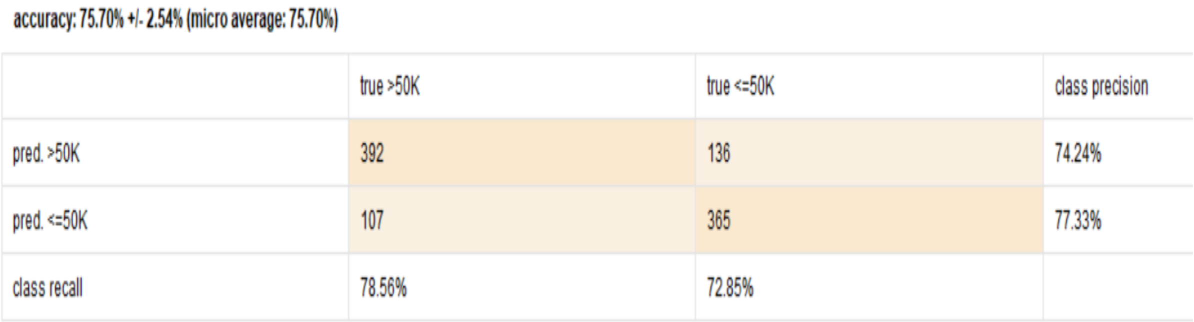
hidden layers	Edit List (2)...
training cycles	200
learning rate	0.1
momentum	0.9
□ *decay*	
☑ *shuffle*	
☑ *normalize*	
error epsilon	1.0E-4
□ *use local random seed*	

Parameters

Neural Net

hidden layers	Edit List (1)...
training cycles	200
learning rate	0.5
momentum	0.0
□ *decay*	
☑ *shuffle*	
☑ *normalize*	
error epsilon	0.001
□ *use local random seed*	

Figure 5 Default ANN Parameters (Left) and Optimizing ANN Parameters (Right)



accuracy: 75.70% +/- 2.54% (micro average: 75.70%)

	true >50K	true <=50K	class precision
pred. >50K	392	136	74.24%
pred. <=50K	107	365	77.33%
class recall	78.56%	72.85%	

Figure 6 The result of Neural Net model (After Adjustment)

Patterns in the data

In conclusion from this testing, the complex algorithm or high hidden layer will decrease the accuracy, which is overfitting. Additionally, the adjusting of Training cycle and Learning rate are the small fraction of impact to the predictive rate. From ANN Layers, “Capital Gain” attribute was the strongest weight in the figure7 comparing to another attributes that means this neuron has the most of impact gradient descent (error adjustment) and important in calculating the final output.

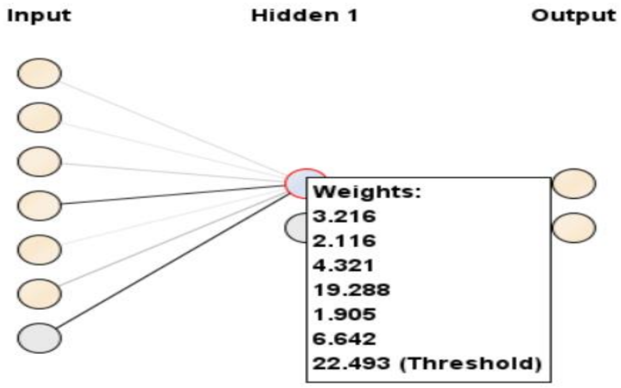
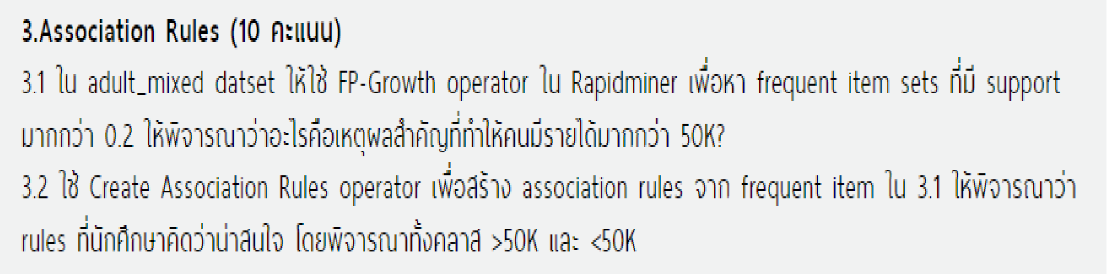


Figure 7 ANN Layers (one node)



3.Association Rules (10 คะแนน)

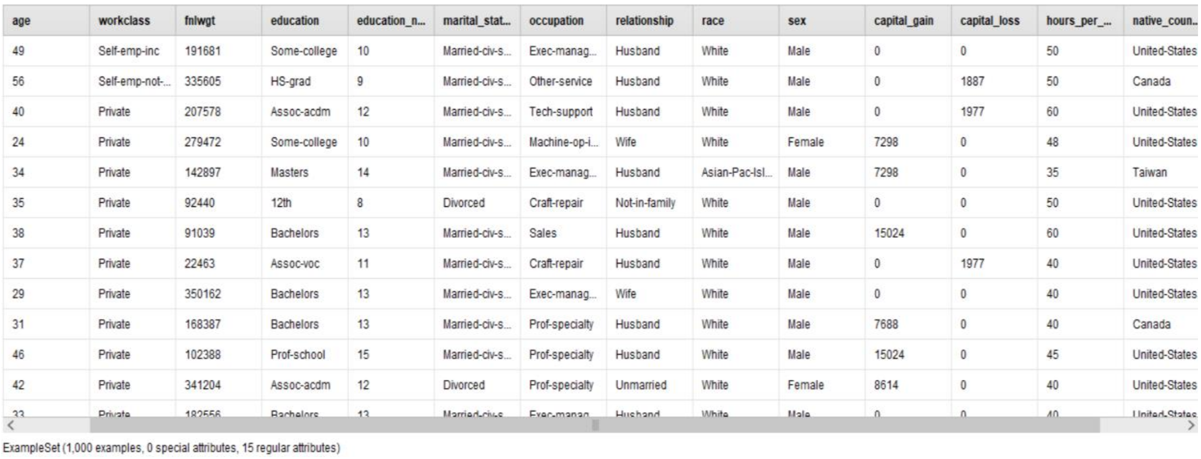
3.1 ใน adult_mixed datset ให้ใช้ FP-Growth operator ใน Rapidminer เพื่อหา frequent item sets ที่มี support มากกว่า 0.2 ให้พิจารณาว่าอะไรคือเหตุผลสำคัญที่ทำให้คนมีรายได้มากกว่า 50K?

3.2 ใช้ Create Association Rules operator เพื่อสร้าง association rules จาก frequent item ใน 3.1 ให้พิจารณาว่า rules ที่นักศึกษาคิดว่าน่าสนใจ โดยพิจารณาทั้งคลาส >50K และ <50K

Part3: Association Rules

Introduction to the dataset and its meta data.

The original data in the figure1 was comprised of 15 attributes, which was contained 1,000 examples, demographically show several aspects such as education, occupation, genders, capital gain and loss, native countries and the range of income between <=50k and >50k.



age	workclass	fnlwgt	education	education_n...	marital_stat...	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_...	native_coun...
49	Self-emp-inc	191681	Some-college	10	Married-civ-s...	Exec-manag...	Husband	White	Male	0	0	50	United-States
56	Self-emp-not-...	335605	HS-grad	9	Married-civ-s...	Other-service	Husband	White	Male	0	1887	50	Canada
40	Private	207578	Assoc-acdm	12	Married-civ-s...	Tech-support	Husband	White	Male	0	1977	60	United-States
24	Private	279472	Some-college	10	Married-civ-s...	Machine-op-i...	Wife	White	Female	7298	0	48	United-States
34	Private	142897	Masters	14	Married-civ-s...	Exec-manag...	Husband	Asian-Pac-Isl...	Male	7298	0	35	Taiwan
35	Private	92440	12th	8	Divorced	Craft-repair	Not-in-family	White	Male	0	0	50	United-States
38	Private	91039	Bachelors	13	Married-civ-s...	Sales	Husband	White	Male	15024	0	60	United-States
37	Private	22463	Assoc-voc	11	Married-civ-s...	Craft-repair	Husband	White	Male	0	1977	40	United-States
29	Private	350162	Bachelors	13	Married-civ-s...	Exec-manag...	Wife	White	Male	0	0	40	United-States
31	Private	168387	Bachelors	13	Married-civ-s...	Prof-specialty	Husband	White	Male	7688	0	40	Canada
46	Private	102388	Prof-school	15	Married-civ-s...	Prof-specialty	Husband	White	Male	15024	0	45	United-States
42	Private	341204	Assoc-acdm	12	Divorced	Prof-specialty	Unmarried	White	Female	8614	0	40	United-States
33	Private	182556	Bachelors	13	Married-civ-s...	Exec-manag...	Husband	White	Male	0	0	40	United-States

ExampleSet (1,000 examples, 0 special attributes, 15 regular attributes)

Figure 1 The original dataset

The dataset was initially explored to understand better about basic characteristics of the data. After the initial exploration, there are some attribute redundance and numeric data such as “sex”, “age”, ”capital gain and loss”, “education numbers”, “fnihgt” and “hours per week” are disposed of the group by Select Attributes operator and using Filter Example operator to sort out the income from “label” attribute in order to answer the 3.1 question. Furthermore, all of nominal data will be transformed into the binominal data by Nominal to Binominal operator. Binominal data is required by FP-Growth operator, which was used to find the frequent items set, set minimum support at 20%. At the end of the model, Create Association Rule operator was included to describe the rule matching and show the vital measurement such as support, confidence, lift and conviction to search for interesting item sets.

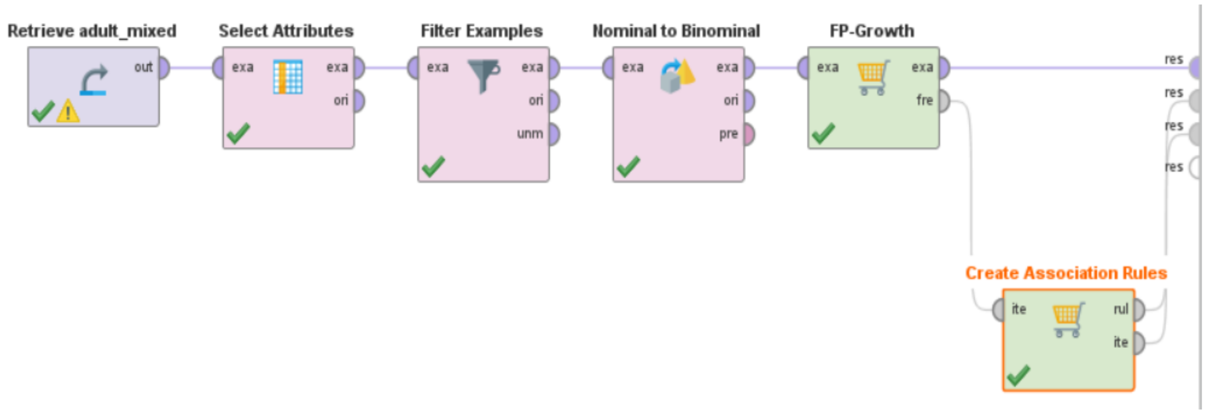
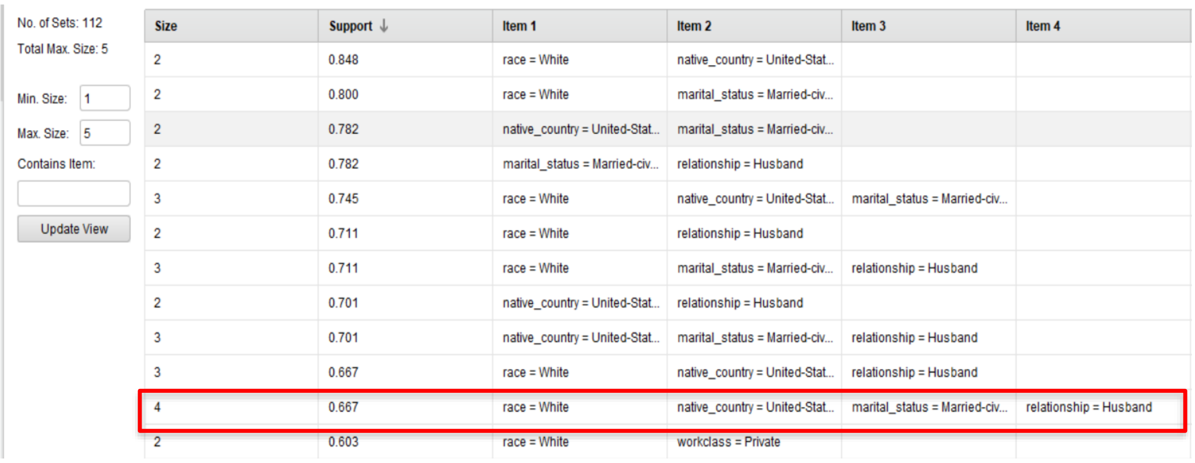


Figure 2 RapidMiner Processes

Frequent Itemset Discussion

The number of examples has an income greater than 50K was 499 transections. The result of frequent items in the figure3 was shown 112 item sets with including 1-itemset to 5-itemset. However, when it was sorted high value of the support, the strong items relationship was demonstrated. For instance, there is one frequent itemset that has 4-itemsets, the support is 0.667, which means 67% of all examples are the people who has income greater than 50K are men, white, married-civ-spouse and was born in the United States. In other words, 334 people from all of examples (499 people) have the 4 same aspects.



No. of Sets: 112
Total Max. Size: 5

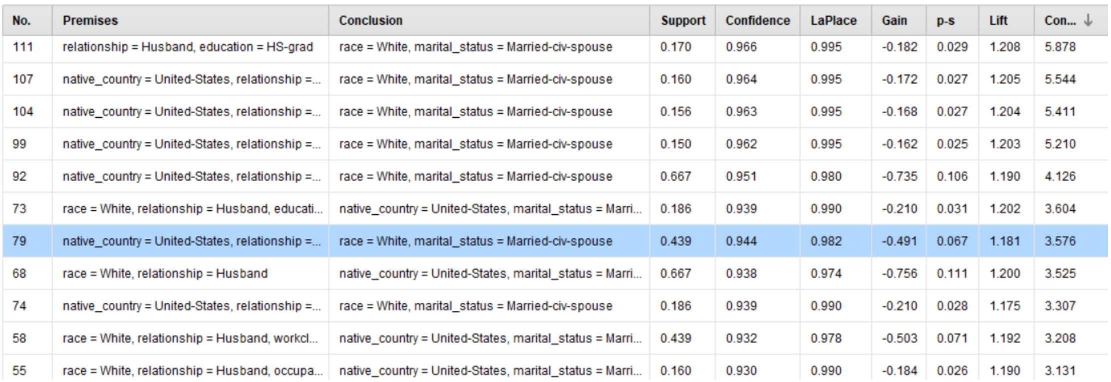
Min. Size: 1
Max. Size: 5
Contains Item:
Update View

Size	Support ↓	Item 1	Item 2	Item 3	Item 4
2	0.848	race = White	native_country = United-Stat...		
2	0.800	race = White	marital_status = Married-civ...		
2	0.782	native_country = United-Stat...	marital_status = Married-civ...		
2	0.782	marital_status = Married-civ...	relationship = Husband		
3	0.745	race = White	native_country = United-Stat...	marital_status = Married-civ...	
2	0.711	race = White	relationship = Husband		
3	0.711	race = White	marital_status = Married-civ...	relationship = Husband	
2	0.701	native_country = United-Stat...	relationship = Husband		
3	0.701	native_country = United-Stat...	marital_status = Married-civ...	relationship = Husband	
3	0.667	race = White	native_country = United-Stat...	relationship = Husband	
4	0.667	race = White	native_country = United-Stat...	marital_status = Married-civ...	relationship = Husband
2	0.603	race = White	workclass = Private		

Figure 3 Frequent item sets (FP-Growth)

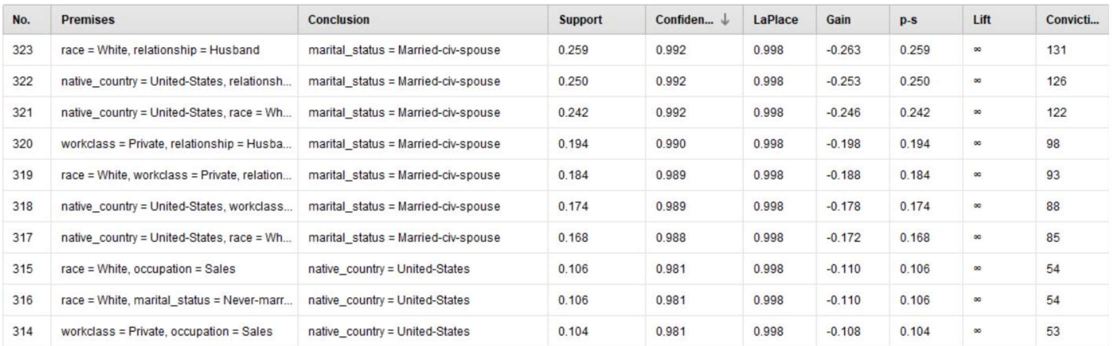
Rules Discussion

The minimum conviction value was set at 1.1 to search for highly depending on the antecede and choosing the high numbers of Confidence, Lift and Conviction. For example in the figure4, people having income >50, the association rule is Native_Country = United-States, Relationship = Husband, Workclass = Private (X) → Race = White, Marital_Status = Married-civ-spouse (Y). Confidence equals 94% means 94% of the transactions that included X and Y. Lift is 1.18 means Confidence (X -> Y) is higher than Support Y, Lift > 1 is a real strength. Confidence is 3.58 means Confidence (X -> Y) is not equal Support Y, so this rule is related. As a result, the high number of Confidence, Lift > 1, Conviction is not equal 1 and value is higher than minimum threshold are all interesting association rule.



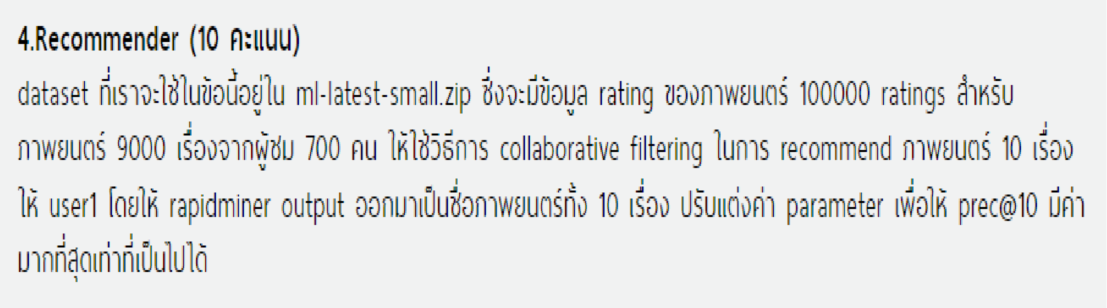
No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Con... ↓
111	relationship = Husband, education = HS-grad	race = White, marital_status = Married-civ-spouse	0.170	0.966	0.995	-0.182	0.029	1.208	5.878
107	native_country = United-States, relationship =...	race = White, marital_status = Married-civ-spouse	0.160	0.964	0.995	-0.172	0.027	1.205	5.544
104	native_country = United-States, relationship =...	race = White, marital_status = Married-civ-spouse	0.156	0.963	0.995	-0.168	0.027	1.204	5.411
99	native_country = United-States, relationship =...	race = White, marital_status = Married-civ-spouse	0.150	0.962	0.995	-0.162	0.025	1.203	5.210
92	native_country = United-States, relationship =...	race = White, marital_status = Married-civ-spouse	0.667	0.951	0.980	-0.735	0.106	1.190	4.126
73	race = White, relationship = Husband, educati...	native_country = United-States, marital_status = Marri...	0.186	0.939	0.990	-0.210	0.031	1.202	3.604
79	native_country = United-States, relationship =...	race = White, marital_status = Married-civ-spouse	0.439	0.944	0.982	-0.491	0.067	1.181	3.576
68	race = White, relationship = Husband	native_country = United-States, marital_status = Marri...	0.667	0.938	0.974	-0.756	0.111	1.200	3.525
74	native_country = United-States, relationship =...	race = White, marital_status = Married-civ-spouse	0.186	0.939	0.990	-0.210	0.028	1.175	3.307
58	race = White, relationship = Husband, workcl...	native_country = United-States, marital_status = Marri...	0.439	0.932	0.978	-0.503	0.071	1.192	3.208
55	race = White, relationship = Husband, occupa...	native_country = United-States, marital_status = Marri...	0.160	0.930	0.990	-0.184	0.026	1.190	3.131

Figure 4 Association Rule of Income > 50K (499 examples)



No.	Premises	Conclusion	Support	Confiden... ↓	LaPlace	Gain	p-s	Lift	Convicti...
323	race = White, relationship = Husband	marital_status = Married-civ-spouse	0.259	0.992	0.998	-0.263	0.259	∞	131
322	native_country = United-States, relationsh...	marital_status = Married-civ-spouse	0.250	0.992	0.998	-0.253	0.250	∞	126
321	native_country = United-States, race = Wh...	marital_status = Married-civ-spouse	0.242	0.992	0.998	-0.246	0.242	∞	122
320	workclass = Private, relationship = Husba...	marital_status = Married-civ-spouse	0.194	0.990	0.998	-0.198	0.194	∞	98
319	race = White, workclass = Private, relation...	marital_status = Married-civ-spouse	0.184	0.989	0.998	-0.188	0.184	∞	93
318	native_country = United-States, workclass...	marital_status = Married-civ-spouse	0.174	0.989	0.998	-0.178	0.174	∞	88
317	native_country = United-States, race = Wh...	marital_status = Married-civ-spouse	0.168	0.988	0.998	-0.172	0.168	∞	85
315	race = White, occupation = Sales	native_country = United-States	0.106	0.981	0.998	-0.110	0.106	∞	54
316	race = White, marital_status = Never-marr...	native_country = United-States	0.106	0.981	0.998	-0.110	0.106	∞	54
314	workclass = Private, occupation = Sales	native_country = United-States	0.104	0.981	0.998	-0.108	0.104	∞	53

Figure 5 Association Rule of Income <= 50K (501 examples)



4.Recommender (10 คะแนน)

dataset ที่เราจะใช้ในข้อนี้อยู่ใน ml-latest-small.zip ซึ่งจะมีข้อมูล rating ของภาพยนตร์ 100000 ratings สำหรับภาพยนตร์ 9000 เรื่องจากผู้ชม 700 คน ให้ใช้วิธีการ collaborative filtering ในการ recommend ภาพยนตร์ 10 เรื่อง ให้ user1 โดยให้ rapidminer output ออกมาเป็นชื่อภาพยนตร์ทั้ง 10 เรื่อง ปรับแต่งค่า parameter เพื่อให้ prec@10 มีค่ามากที่สุดเท่าที่เป็นไปได้

Part 4: Recommendation Engine

Introduction to the dataset and its meta data.

The rating dataset was contained 100,836 rating transections for 9,742 movies from 610 users. There are 4 attributes that comprise of “User_ID”, “Movie_ID”, “Rating” was given 1 to 5 and “Timestamp”. Moreover, the movies dataset is another information that was used for describing movies name, consisting of 3 attributes such as “Movie_ID”, “Title” and Genres.

Recommendation algorithms used

First of all, the rating dataset was linked with Set Role operator to identify users and items. And then, Split data operator was used for dividing the training dataset into 80% of all information and the testing dataset was separated into 20%. Next steps, the training dataset was trained by User K-NN operator from collaborative filtering item recommendation and sent to Apply model (item recommendation) operator while receive testing dataset from Split Data operator. Finally, all of testing data will be applied with the recommendation model and Join operator was used to match up the coincident ID attribute between “Item_ID” attribute from the consequence of recommendation model and “Movie_ID” attribute from the movies dataset to show titles and genres movies.

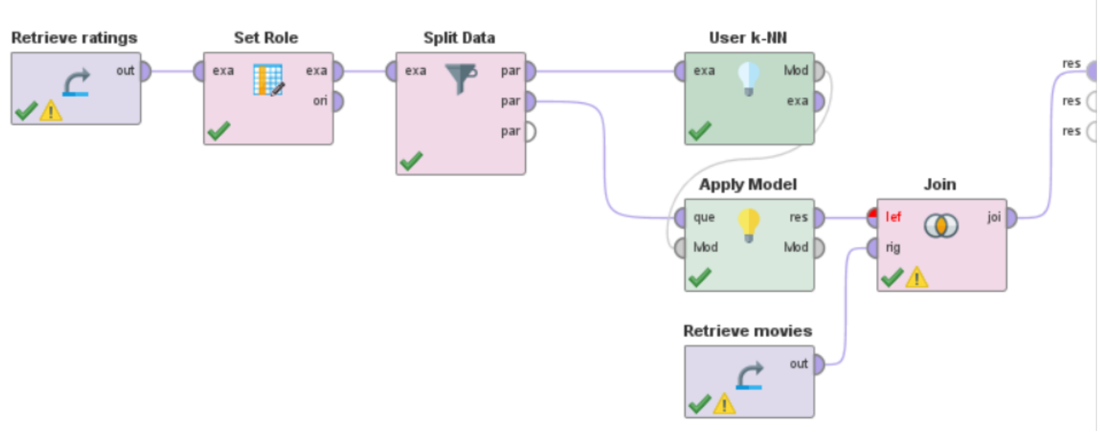


Figure 1 The model of Recommending 10 Movies for User1

For assessing accuracy, Performance (Item recommendation) operator was used to evaluate the accuracy of movies recommendation in the figure2.

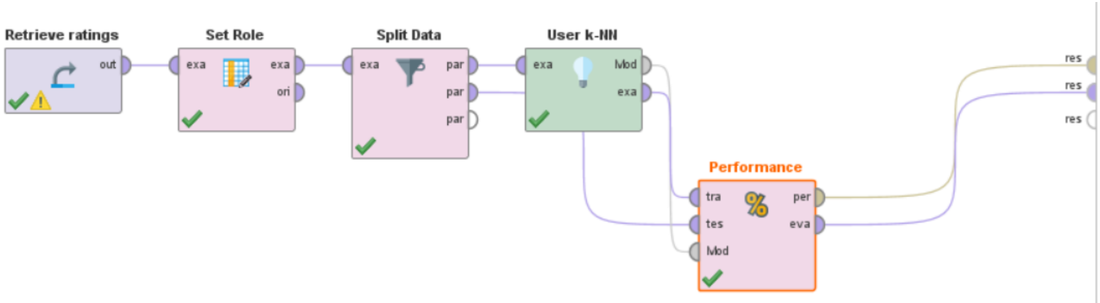
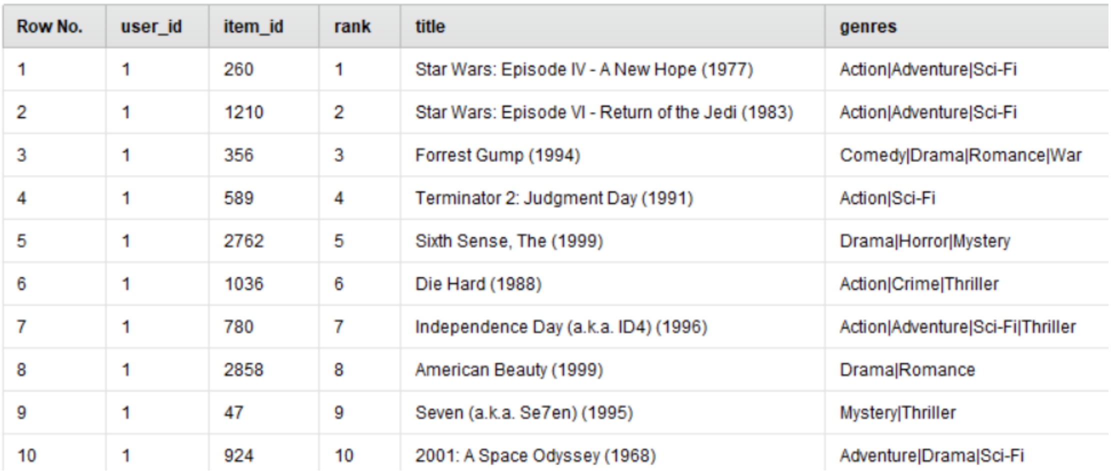


Figure 2 The model of evaluating accuracy of recommendation

Results of training using default parameters.

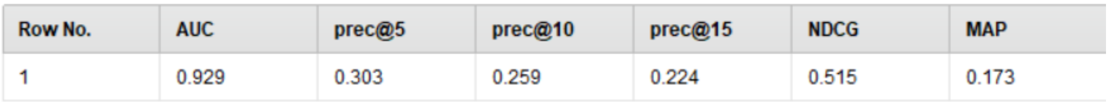
The consequence of this model will show the recommendation of 10 movies ranking for User 1 in the figure3. Whereby, the recommendation model predicts the rating movies and occupy the predictive rating in every movie. And, this model will recommend the ranking of unseen movie relating to the high score of predictive movies that has similar genres to the favorite movie for each user. For example, Action movie was predicted the highest rating for User1. Therefore, the kind of unseen movies are similar type of action movie and predictive high score will be advised to User1 such as Star Wars Episode 4 and 6, Terminator 2, Die Hard and Independence Day.



Row No.	user_id	item_id	rank	title	genres
1	1	260	1	Star Wars: Episode IV - A New Hope (1977)	Action|Adventure|Sci-Fi
2	1	1210	2	Star Wars: Episode VI - Return of the Jedi (1983)	Action|Adventure|Sci-Fi
3	1	356	3	Forrest Gump (1994)	Comedy|Drama|Romance|War
4	1	589	4	Terminator 2: Judgment Day (1991)	Action|Sci-Fi
5	1	2762	5	Sixth Sense, The (1999)	Drama|Horror|Mystery
6	1	1036	6	Die Hard (1988)	Action|Crime|Thriller
7	1	780	7	Independence Day (a.k.a. ID4) (1996)	Action|Adventure|Sci-Fi|Thriller
8	1	2858	8	American Beauty (1999)	Drama|Romance
9	1	47	9	Seven (a.k.a. Se7en) (1995)	Mystery|Thriller
10	1	924	10	2001: A Space Odyssey (1968)	Adventure|Drama|Sci-Fi

Figure 3 The result of 10 movies recommendation for the User1.

The result of evaluating accuracy for prec@10 is 25.9%, which means the Percentage of 10 movies in a recommendation list that the user would rate as useful. In short, the successful recommendation is 25.9% that means 2.59 movies have higher predictive rate than the average calculation over all users or useful advice from top 10 recommended.



Row No.	AUC	prec@5	prec@10	prec@15	NDCG	MAP
1	0.929	0.303	0.259	0.224	0.515	0.173

Figure 4 The rank accuracy of default recommendation (K = 80)

Results when training using modified parameter settings.

To improve the accuracy of this model, K parameter is number of nearest neighbors, which are randomly tested in order to improve the effective accuracy. As one can see K = 40 is optimal number from all of testing that has the highest accuracy comparing to other numbers. K = 40 means the 40 number of movies, which have data points close to favorite movie are considered to be rating and recommend unseen movies to the users.

Prec@k	**K =10**	**K=40** **(Optimization)**	**K= 50**	**K=80** **(Default)**	**K=100**	**K=1,000**
prec@5	27.5%	31.8%	31.2%	30.3%	30.5%	19.0%
prec@10	24.1%	**26.4%**	26.1%	25.9%	25.2%	15.5%
pre@15	20.8%	23.1%	22.9%	22.4%	22.2%	14.2%

Figure 5 The rank accuracy of default recommendation (k = 80)

Conclusion

Item-based collaborative filtering is suitable model to efficiently recommend unseen movie to the clients due to the number of users dominates the number of movies in the system. Therefore, the recommendation model will predict the rating movie involving with similar movie by using K nearest neighbor numbers to be considered the number of similar movies. And then, those similar movies are calculated to occupy the rating. Eventually, the unseen movie will be raked and recommend to each user. As a result of adjusting K numbers, prec@10 is improved an accuracy from 25.9% to 26.4% by changing number of k from 80 to 40.