Клъстерен анализ на резултати от анкета за изследване удовлетвореността от обучението по Компютърна математика

Стефка Караколева

Abstract: Cluster Analysis of the Results of a Poll about the Satisfaction with the Education in Computer Mathematics. *In this paper the results from a research about the satisfaction of computer-based education in Mathematics are presented. A cluster analysis of the results from a poll is done by means of SPSS Hierarchical Cluster Analysis, Non Hierarchical Cluster Analysis and Two Step Cluster Analysis. The derived from SPSS results are discussed.*

Key words: *computer-based education, cluster analysis, statistics, mathematics, SPSS, MATLAB*

ВЪВЕДЕНИЕ

Чрез програмната система SPSS е извършен клъстерен анализ на резултатите от анкета [5] за изследване удовлетвореността от обучението по Компютърна математика, проведена в периода 2012-2014 г. Общият брой на изследваните лица е 151, като част от тях са анкетирани чрез разработения on-line вариант, публикуван на адрес: http://landing.zlatarov.info/polls/index.php/338631/lang-bg

Изследванията, предмет на настоящата статия, са част от задълбочено проучване на резултатите от прилагането на компютърно съпроводено обучение на студенти – бъдещи инженери, обучавани в Русенски Университет [2, 3, 4].

ИЗЛОЖЕНИЕ

Целта на клъстерния анализ е n на брой обекти (анкетирани) да се групират в k групи (клъстери) като се използват p на брой променливи (признаци) [1,6,7,8,9].

Хипотезата при изследването е, че по отношение оценката на идеята за компютърно съпроводено обучение и ползата от него, както и намеренията на анкетираните да използват система за математически изчисления в бъдеще, изследваните случаи се оформят в групи (клъстери).

В качеството на изследвани признаци се използват: „1-Оценете по скала от 2 до 6 идеята часовете по математика да се провеждат в компютърна зала с използване на CAS", „2-Ще продължите ли да използвате MATLAB за решаване на математически задачи?“, „3-Според вас, използването на CAS ще спомогне ли за повишаване нивото на обучение по математика?“ и „4-Има ли бъдеще компютърно съпроводеното обучение по математика?“.

Чрез вариационен анализ се установява разпределението на анкетираните според отговорите им по изследваните въпроси 1-4. Установено е, че голяма част от анкетираните подкрепят идеята за компютърно съпроводено обучение по математика. Предполага, че е възможно те да се групират в клъстери, чийто брой се установява в процеса на клъстерния анализ.

При изследването са използвани трите вида методи за клъстерен анализ [1,7,9]:

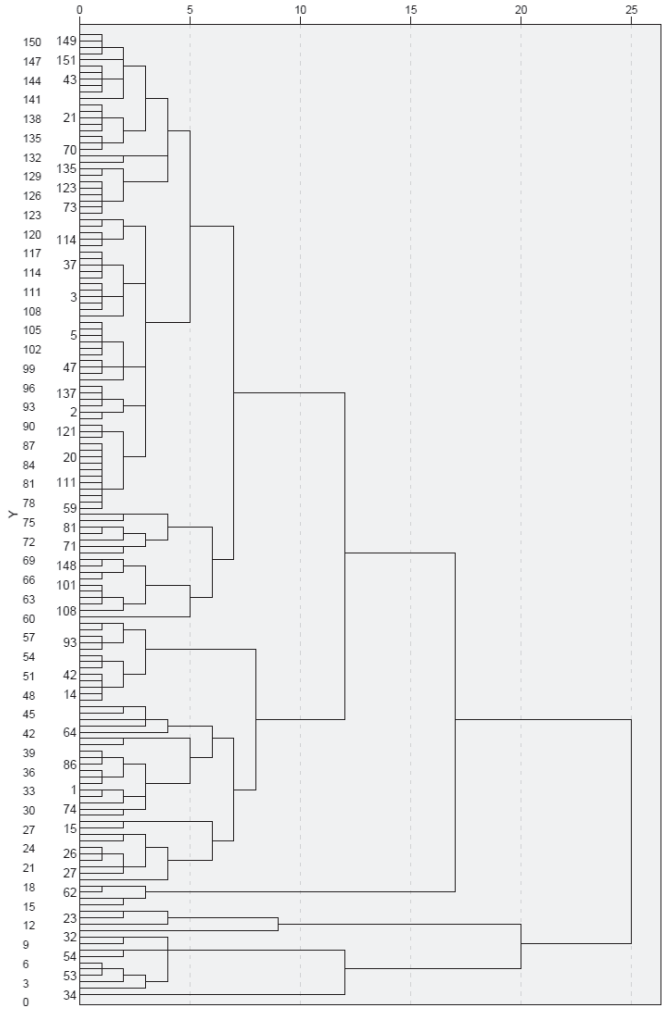
- Йерархична клъстеризация (Hierarchical Cluster Analysis);
- Нейерархична клъстеризация (Non Hierarchical Cluster Analysis);
- Двустъпкова клъстеризация (Two Step Cluster Analysis).

Извършеният анализ е съчетание между йерархична и нейерархична клъстеризация. Йерархична клъстеризация се използва за определяне броя на клъстерите. Най-напред се прилагат всички възможни подметоди за йерархична клъстеризация, преценява се кога се получава добро решение, записват се резултатите от добрите решения и се сравняват. След това се прилага нейерархична клъстеризация и се записват резултатите. Сравняват се резултатите

от йерархична и нейерархична клъстеризация и се определя кой метод е най-подходящ. Накрая се прави двустъпкова клъстеризация, сравняват се получените променливи от най-подходящите методи и се извеждат резултатите.

1. ЙЕРАРХИЧНА КЛЪСТЕРИЗАЦИЯ

Дендрограма. Основен инструмент при йерархичната клъстеризация е *дендрограмата* — графично изображение, което показва процеса на формиране на клъстерите чрез присъединяване на случаите към тях. При това, поради наличните в SPSS седем метода за йерархична клъстеризация, е направено изследване с всеки един от тях, след което получените дендрограми се анализират и се избира най-подходящия метод и броя на клъстерите. Основен критерий при анализа на дендрограмите е да се формират малък на брой клъстери с възможно най-голям брой присъединени елементи.



Фигура 1. Дендрограма - йерархична клъстеризация по метода на междугрупово свързване

На фиг. 1 е изобразена дендрограмата, получена при използване на метода на междугрупово свързване. За конкретното изследване се установява наличието на един голям и два по-малки клъстери, които биха могли да се обединят в един по-голям клъстер.

1.1. Определяне на подходящ брой клъстери и метод на клъстеризация

Определянето на най-подходящия брой клъстери е извършено чрез прилагане на всички подметоди на йерархична клъстеризация:

- метод на междугрупова свързаност (Between-groups linkage);
- метод на вътрешногрупова свързаност (Within-groups linkage);
- метод на най-близкия съсед (Nearest neighbor);
- метод на най-далечния съсед (Furthest neighbor);
- центроиден метод (Centroid clustering);
- медианен метод (Median clustering);
- метод на Вард (Ward's method)

и нейерархична клъстеризация по метода на K-средни (K-means clustering).

За всеки метод са извършени изчисления при брой на клъстерите от 2 до 4 като се записват получените резултати в нови променливи, съдържащи цели числа от 1 до k, $k = 2, 3, 4$, които показват за всеки анкетиран към кой номер клъстер е присъединен. След вариационен анализ на генерираните от съответните методи променливи, са получени данни за разпределението на случаите за всеки отделен метод при $k = 2, 3, 4$. За определяне на най-подходящия метод за клъстеризация, е приложен сравнителен анализ на получените променливи за всеки отделен метод при фиксиран брой клъстери $k = 2,3,4,$ като се търси най-доброто съответствие между тях чрез коефициента на контингенция (Contingency Coefficient).

Анализът на получените резултати показва, че:

- При формиране на два клъстера $k = 2$, нейерархичната клъстеризация дава по-лоши резултати в сравнение с методите на йерархична клъстеризация. От всички методи, най-добро съответствие по отношение принадлежност на анкетираните към клъстерите се получава за методите Between-groups clustering и Ward's method, което личи от най-високата стойност на коефициента на контингенция (0.693). По метода на най-близкия съсед във втори клъстер попада само един случай, което категорично показва, че методът не е подходящ (коефициенти на контингенция под 0.2);
- При формиране на три клъстера $k = 3$, най-добро съответствие между променливите, показващи принадлежност към трите клъстера има при метода на междугрупово свързване и при центроиден метод - най-висок коефициент на контингенция 0.764;
- При формиране на четири клъстера $k = 4$, най-добро съответствие между променливите, показващи принадлежност към четирите клъстера има отново при методите на междугрупово свързване и центроиден метод - най-висок коефициент на контингенция 0.819.

Основният извод от комбинираното прилагане на йерархична и нейерархична клъстеризация е, че е удачно да се изберат *два* на брой клъстери, тъй като при деление на 3 клъстера повечето случаи са разпределени в 2 от тях, а в третия попадат малък брой единици. По отношение на метода, най-подходящ се оказва *методът на междугрупово свързване*, който е с най-добри показатели при всички сравнения.

1.2. Определяне на съдържателния смисъл на клъстерите и приноса на признаците при формиране на клъстерите

След определяне броя на клъстерите и най-ефективния метод на клъстеризация, възникват логично два въпроса:

- Какъв е профилът на отделните случаи, попадащи във всеки един от клъстерите;
- Доколко (в каква степен) всеки един от разглежданите признаци (въпроси) допринася за разделянето на случаите в клъстери.

Дисперсионен анализ. За да се отговори на поставените въпроси и да се провери в каква степен отделните признаци разделят случаите в клъстери, се извършва *дисперсионен анализ.* Като независими променливи се въвеждат четирите признаци (въпроси), използвани при клъстеризацията, а като зависими променливи - получените след клъстеризация променливи, показващи принадлежността на всеки отделен случай към клъстерите.

Анализът е извършен за всички методи на йерархична и нейерархична клъстеризация. В таблици 1 и 2 са дадени резултатите от дисперсионния анализ за четирите признака и зависимата променлива, получена по метода на междугрупово свързване, показваща принадлежността на случаите към двата клъстера. Подобни резултати се получават и при анализа на променливите, показващи принадлежност към клъстер, получени по останалите методи.

Таблица 1.

Резултати от дисперсионен анализ - метод на междугрупово свързване

ANOVA Table

		Sum of Squares	df	Mean Square	F	Sig.
1-Оценете по скалата от 2 до 6 идеята часовете по математика да се провеждат в компютърна зала с CAS * Average Linkage (Between Groups)	Between Groups	,173	1	,173	,177	,675
	Within Groups	146,131	149	,981		
	Total	146,305	150			
2-Ще продължите ли да използвате Matlab за решаване на математически задачи * Average Linkage (Between Groups)	Between Groups	8,104	1	8,104	6,314	,013
	Within Groups	191,247	149	1,284		
	Total	199,351	150			
3-Според вас, използването на CAS ще спомогне ли за повишаване на качеството на обучение по математика? * Average Linkage (Between Groups)	Between Groups	17,294	1	17,294	45,4	,000
	Within Groups	56,799	149	,381		
	Total	74,093	150			
4-Има ли бъдеще компютърно съпроводеното обучение по математика? * Average Linkage (Between Groups)	Between Groups	124,012	1	124,01	322	,000
	Within Groups	57,419	149	,385		
	Total	181,430	150			

От анализа на резултатите от дисперсионния анализ се оформят няколко важни извода:

- Не се установява зависимост между отговорите на анкетираните по първия въпрос и принадлежността им към клъстерите. Този факт се установява чрез сравняване на средните стойности на признака за двата клъстера. И в двата случая, средната стойност на отговорите по този въпрос е близка до средната стойност на цялата извадка. Този извод се потвърждава и от стойността на равнището на значимост за първи въпрос. Стойността на равнището на значимост $Sig = 0.675 > 0.05$ (таблица 1) надвишава риска за грешка, което означава, че се приема нулевата хипотеза, т.е. няма статистически значима връзка между отговорите на анкетираните по 1. въпрос и принадлежността им към определен клъстер. Таблица 2 носи информация и за коефициентите на корелационно отношение и определеност. Анализът на коефициента на определеност $Eta\ squared = 0.001$ показва, че едва 0.1 % от различията в принадлежността към определен клъстер се дължат на отговорите на 1. въпрос. Следователно, първият въпрос може да отпадне при окончателното клъстеризиране, тъй като той има незначителен принос при формиране на клъстерите;

Таблица 2.
Стойности на корелационно отношение и коефициент на определеност при дисперсионен анализ

Measures of Association

	Eta	Eta Squared
1-Оценете по скалата от 2 до 6 идеята часовете по математика да се провеждат в компютърна зала с CAS * Average Linkage (Between Groups)	,034	,001
2-Ще продължите ли да използвате Matlab за решаване на математически задачи * Average Linkage (Between Groups)	,202	,041
3-Според вас, използването на CAS ще спомогне ли за повишаване на качеството на обучение по математика? * Average Linkage (Between Groups)	,483	,233
4-Има ли бъдеще компютърно съпроводеното обучение по математика? * Average Linkage (Between Groups)	,827	,684

- По отношение на останалите въпроси, се наблюдава значима връзка между признаците и принадлежността към определен клъстер. Това личи от таблица 1, където за 2, 3 и 4. въпроси стойностите в последната колона на равнището на значимост са по-малки от риска за грешка 0.05. От таблица 2, въз основа на стойностите на коефициента на определеност може да се каже, че 4.1% от различията в принадлежността към определен клъстер се дължат на отговорите по 2. въпрос; съответно за 3. въпрос - приносът му при определяне на клъстерите е 23.3% и с най-голям принос - 68.4% е четвъртият въпрос.

След дисперсионния анализ може да се даде и съдържателна характеристика на това, какви случаи попадат съответно в 1. и 2. клъстер.

Сравнителният анализ на средните стойности на отговорите на въпросите показва, че в първия клъстер (137 случая) попадат хора, които дават високи оценки по всички въпроси - средната стойност по 1. въпрос е 4.88 при 4.89 за цялата извадка; за 2. въпрос средната стойност е 3.94 при 3.87 за извадката, за 3. въпрос - съответно 2.45 при 2.34 за извадката и за 4. въпрос - 4.48 при 4.19 за извадката.

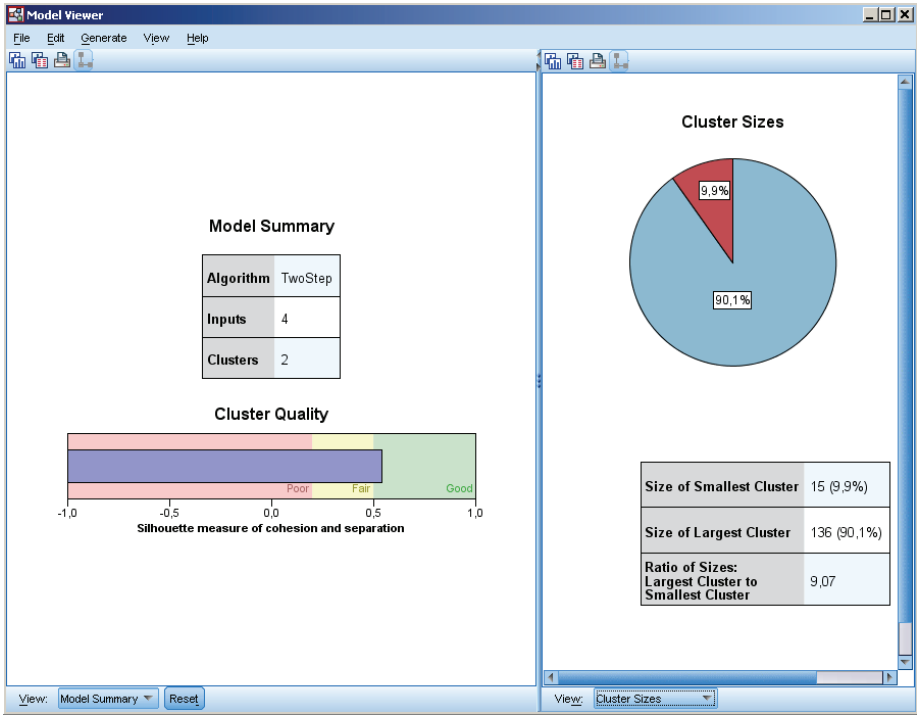
Във 2. клъстер, в който има 14 случая, попадат хора, дали ниски оценки по 3 от въпросите и оценка близка до средната (дори по-висока от средната) по първи въпрос. За 2. въпрос средната стойност е 3.14 при 3.87 за извадката, за 3. въпрос - 1.29 при 3.24 за извадката и за 4. въпрос - 1.36 при 4.19 за извадката.

Можем условно да наречем 1. клъстер „Клъстер на положителна оценка за Компютърната математика“ и 2. клъстер „Клъстер на отрицателна оценка на КМ“.

2. Двустъпкова клъстеризация

След като са определени броят на клъстерите, броят единици в клъстерите и профилът на случаите, попадащи в клъстерите, се прилага методът на *двустъпкова клъстеризация* за проверка и графично изобразяване на получените резултати.

На фигура 2 е изобразен графичен прозорец с резултатите от двустъпковия клъстерен анализ на извадката с отговорите на анкетата по четирите въпроса.



Фигура 2. Графично изобразяване на модел с два клъстера по метода на двустъпкова клъстеризация

Полезен инструмент за визуална оценка качеството на клъстеризация по този метод е *измерителят за близост и различие* (Silhouette measure of cohesion and separation), позициониран в трицветна скала. Ако измерителят е в червената зона, резултатът е лош, ако е в жълтата зона - задоволителен и ако е в зелената - добър.

Получените резултати по метода на двустъпкова клъстеризация са сходни с тези, получени по метода на междугрупово свързване и метод на Ward - получават се автоматично два клъстера съответно със 136 и 15 случая, като измерителят на близост и различие показва добро качество на модела, тъй като е позициониран в

зелената зона, фигура 2. В дясната част на графичния прозорец е показана кръгова диаграма, изобразяваща размерите на клъстерите, таблица с характеристиките на най-големия и най-малкия клъстер, както и съотношението между тях (9:1).

Таблица 3.
Сравнение между методите на двустъпкова клъстеризация и междугрупово свързване

		Average Linkage (Between Groups)		
		1	2	Total
TwoStep Cluster Number	1	135	1	136
	2	2	13	15
Total		137	14	151

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Contingency Coefficient	,663	,000
N of Valid Cases		151	

Извлечена е информация от графичния прозорец на фигура 2 за признаците, подредени по степен на важност при формиране на клъстерите. От получените резултати се потвърждава заключението, че 1. въпрос не влияе съществено при формиране на клъстерите - неговият принос е 0.01 от най-важния 4. въпрос, чиято мярка е единица. Извършена е двустъпкова клъстеризация при фиксирани три на брой клъстери, която се оказва с по-лошо качество, отколкото при два клъстера - измерителят на близост и различие е в жълтата (средна) зона.

3. Сравнителен анализ

След прилагането на всички методи за клъстерен анализ, се прави сравнение на съответствието между променливите, получени по методите на междугрупово свързване и двустъпкова клъстеризация чрез сравнителен анализ на съответните променливи, таблица 3. Полученият значим коефициент на контингенция 0.663 показва добро съответствие между променливите.

ИЗВОДИ

В резултат на извършения клъстерен анализ на резултатите от проведената анкета по Компютърна математика е установено, че всички анкетирани се групират в два клъстера според тяхната оценка за ползата от компютърната математика и намерението им да я използват в бъдеще. При това над 90% от анкетираните оформят клъстера на „положително оценяващи компютърната математика“. Резултатите от клъстерния анализ доказват удовлетвореност и подкрепа на новата методика на компютърно съпроводеното обучение по математика.

ЛИТЕРАТУРА

[1] Гочева-Илиева, С. Г. *Вероятности и статистика*, Университетско издателство „Паисий Хилендарски“, Пловдив, 2013.

[2] Караколева,С. *Дискриминантен анализ на резултатите от обучението по Висша математика в Русенски Университет*, Научни трудове на РУ&СУ, том 54, серия 6.1, Русе, 2015.

[3] Караколева, С. *Изследване резултатите от обучението по Висша математика в Русенски Университет чрез класификационни дървета,* Научни трудове на РУ&СУ, том 54, серия 6.1, Русе, 2015.

[4] Караколева, С. *Доказване ефективността на компютърно съпроводено обучение по Висша математика в Русенски Университет,* Научни трудове на РУ& СУ, том 54, серия 6.1, Русе, 2015.

[5] Караколева, С. *Изследване удовлетвореността от обучението по*

[6] *Компютърна математика,* Научни трудове на РУ&СУ, том 53, серия 6.1, стр. 46-51, Русе, 2015.

[7] Крыштановский, А.О, *Анализ социологических данных с помощю пакета SPSS*, Издательский дом ГУ ВШЭ, Москва, 2006.

[8] Моосмюлер, Г., Ребик, Н. *Маркетинговые исследования с SPSS*, ИНФРА-М, Москва, 2009, ISBN: 978-5-16-002811-8.

[9] Наследов, А. *SPSS - компютерный анализ данных в психологии и социальных науках*, Питер, Москва, 2005, ISBN: 5-318-00703-1.

[10] Харалампиев, К. *Въведение в основните статистически методи за анализ*, ИК „Йозеф Кнехт“, София, 2007, ISBN: 978-954-07-2847-6.

За контакти:

Стефка Караколева, Катедра “Приложна математика и статистика”, Русенски университет “Ангел Кънчев”, тел: 082-888 606, e-mail: skarakoleva@uni-ruse.bg