Оценяване възможностите за диагностициране на заболели от Фузариоза царевични семена чрез спектрален анализ и SIMCA метод

Виолета Манчева

Assess the ability to diagnose infected with Fusarium corn seeds by spectral analysis and SIMCA method: *An approach for assessing the spectral characteristics of healthy and infected with Fusarium corn seeds of seven varieties by SIMCA method is presented. A SIMCA model was created with data from the training set. Data processing was carried out with three kinds of transformations: smooth, first and second derivatives. As a result of this ten factors of significance in the classification of seeds were evaluated for each of these three processes. It was found that only the first factor is significant when processing data with smoothing, while data processing with first and second derivatives - only first three factors are significant. A total percentage of correctly and incorrectly identified corn seeds is calculated from a test set of three types of transformations. The results show that the highest percentage of correctly identified healthy seeds (90.36%) obtain in the model with the first derivative, and proper recognition infected seeds (96.43%) - in the model with the second derivative.*

Key words: *corn seeds, spectral characteristics, Fusarium, SIMCA method.*

ВЪВЕДЕНИЕ

Царевицата е една от основните зърнени култури в света. Качество на царевични семена се определя от редица показатели за качество, регламентирани в българските стандарти [9]. Един от основните показатели за качеството е болестта Fusarium. Това заболяване е най-значително в икономическо отношение и е токсично както за хората, така и за животните. Това е причината да се работи интензивно в областта на разпознаването на тази болест.

Методите за определяне на качествените показатели най-общо могат да се разделят на методи, оценяващи вътрешни качествени показатели и методи, оценяващи външни качествени показатели на обектите. Вътрешните показатели дават информация за съдържанието на влага в семето, вътрешни дефекти, твърдост и състав.

При използване на визуални методи се анализира само повърхността на продукта или неговата видима част. За да се оцени състоянието на вътрешната структура се използват методи, базирани на спектрален анализ на различни участъци на електромагнитния спектър. При анализиране на спектралните характеристики във вълновия диапазон на видимата (VIS) и близката инфрачервена (NIR) области могат да се оценят различни качествени показатели, като степен на зрялост, съдържание на сухи вещества, **заболяване**, цвят ,влага, захари и др.

В областта на Image Processing са направени изследвания за диагностициране на заболяването Фузариоза по царевични семена чрез анализ на цифрови изображения. Създадени са методики за диагностициране на здрави и заразени царевични семена по цветови признаци за няколко цветови модела. В областта на разпознаване на образи е направена класификация на царевични семена чрез невронни мрежи и Fuzzy логика. Направените изследвания дават добри резултати, но те са получени на базата на външните качествени показатели на царевичните семена [10].

Резултатите от диагностицирането на заболяването Фузариоза, получени на базта на външни качествени показатели, не гарантират, че вътрешната структура на семената е напълно здрава. Това налага да се прави и спектрален анализ на изследваните семена.

Разработени са алгоритъми за разпознаване на заболели от Фузариоза царевични семена чрез анализ на спектралните им характеристики на базата на тяхното описание с линейни дискретни модели [3, 11]. В [11] е разработен алгоритъм

за разпознаване на болни семена. Тъй като, от една страна са изследвани спектрите на малка извадка семена, а от друга – използвания спектрофотометър не е толкова прецизен, то този алгоритъм е усъвършенстван и повторен в [3]. Получените резултати показват, че само за четири от изследваните седем сорта този начин на класификация е надежден. Разработен е и подход за разпознаване на същото заболяване, базиран на обработката на спектралните данни и Wavelet трансформация [4]. Получените резултати от двата класификатора са различни – с линейния е получено 100% разпознаване на заразените семена, а с вероятностната невронна мрежа – 93,3%.

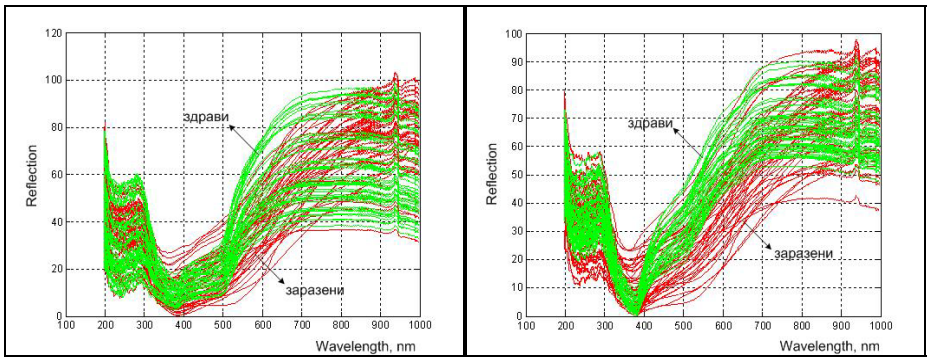
С голяма популярност се използва метода Soft Independent Modeling of Class Analogy (**SIMCA**) за анализ на спектрални данни. Той дава оценка за класификацията на обекти от няколко класа [1, 2, 6, 8]. Тъй като до сега не е правена оценка на царевични семена с този метод, то би било интересно да се провери дали той ще даде надеждна класификация.

Целта на доклада е оценка на спектрални характеристики на здрави и заразени с Фузариоза царевични семена чрез SIMCA метод.

ИЗЛОЖЕНИЕ

1.Обект на изследването

Обект на изследването е заболяването розова Фузариоза (*Fusarium Moniliforme*) и проявата му по царевичните семена. Изследвани са извадки от седем сорта семена – Кнежа 308, Кнежа 436, Кнежа 613, Кнежа 620, 26А, ХМ87/136 и Русе 424. Спектралните им характеристики са получени чрез спектрофотометър на фирмата Ocean Optics във видимата и близката инфрачервена област в спектралния диапазон от 200 до 1000 nm. За всеки от сортовете са снети спектралните характеристики на дифузно отражение на 50 здрави и 50 заразени семена за двете им страни – зародиш и гръб. На фиг.1 са показани получените спектрални характеристики за един сорт. За останалите шест сорта те изглеждат по подобен начин.



а) гръбна страна *б) страна на зародиш*

*Фиг. 1 Спектрални характеристики на дифузно отражение за 50 здрави и 50 заразени царевични семена от **сорт Кнежа 308***

2. Методика на експеримента

Методите за класификация се базират на няколко подхода – разделяне, вероятност и подобие. Подходът, свързан с подобие допуска, че подобните проби се намират в една и съща част от пространството, определено от измерваните параметри. Към тази група спада метода SIMCA. Той е двуетапен. Първо се създава

модел на базата на проби от групата за калибровка и след това този модел се използва за класификация на неизвестни проби.

Методиката за класификация на царевичните семена със SIMCA метод включва следните основни стъпки:

1) Първата стъпка е изследваните обекти да се разделят на класове.

За целта царевичните семена са разделени на два класа – клас 1 (здрави) и клас 2 (заразени).

2) Втората стъпка е разделянето на пробите на две групи – за обучение (калибровка) и за тестване.

За всеки от двата класа семената са разделени в двете групи (обучение и тест) по метода на Кеннард и Стоун [5]. При него се задава желания брой обекти, които да бъдат подбрани от изходните. В оригиналния алгоритъм като метрика се използва евклидовото разстояние. Обучаващата група включва 30 спектъра на здрави и 30 спектъра на заразени царевични семена за всеки от седемте сорта. Тестовата група включва 20 спектъра на здрави и 20 спектъра на заразени царевични семена за всеки сорт.

3) Третата стъпка включва създаване на независими модели.

SIMCA разработва модели на базата на главни компоненти (PC) за всяка обучаваща категория. Качеството на метода Principal Component Analysis (PCA) да определя главните направления на представяното множество данни, позволява трансформиране на данните в собственото пространство и по този начин да намалява размерността на решаваните задачи. Така съхраняването на данните е по-ефективно, а и търсенето на информация в едно редуцирано пространство става по-бързо.

При създаването на SIMCA модела за всеки сорт се използват данните от обучаващата група.

Методът SIMCA дава възможност да се избере начина на преобразуване на данните и броя на факторите, които участват при създаването на модела за класа. В конкретния случай са избрани броя на тези фактори да бъде 10.

При снемането на характеристиките със спектрофотометъра се получават шумове в сигналния поток. Ефектът от тези случайни изменения се намалява чрез изглаждане на характеристиките. Също така, спектрите могат да съдържа едва доловими стъпаловидни върхове, които са от важно значение за търсенето на разлики между двата класа. За откриването на тези върхове се изчислят първа и втора производни на спектърите за подобряване на резултатите. Преобразуванията с първа и втора производна, както и изглаждането са базират на полиномния филтър от втори ред на Savitzky-Golay [7].

4) Четвъртата стъпка включва обработка на данните със SIMCA модела.

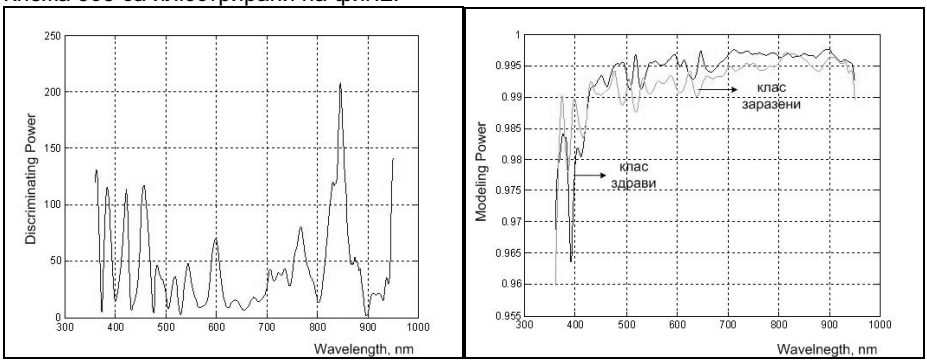
Обработката на данните е извършена чрез използване на трите трансформации, споменати по-горе. В резултат на всяка една от тези три обработки се получават няколко параметъра, които са от важно значение при класификацията на семената.

Параметърът **Interclass Distance** показва разстоянията между двата класа (клас „здрави" и клас „заразени") в модела. Колкото това разстояние е по-голямо, толкова разделимостта между класовете е по-добра.

Параметърът **Eigenvalues** дава в проценти стойностите на факторите, от които е съставен модела за съответния клас. Факторите, които имат стойности близки до 100%, са най-значими за разделянето на класовете.

Параметърът **Class Distances** дава разстоянията от всяка проба до центровете на моделите за класовете. Тези разстояния се представят чрез множество диаграми, чиито брой зависи от броя на използваните класове. Двойките комбинации от класовете формират осите на диаграмата.

При обработката на данните от обучаващия модел се получават два показателя – **Discriminating Power** и **Modeling Power**, графиките на които за сорт Кнежа 308 са илюстрирани на фиг.2.



а) б)

*Фиг. 2 Discriminating Power (а) и Modeling Power (б) за сорт **Кнежа 308** при изглаждане на спектралните характеристики*

Дължините на вълните, които са отговорни за класификацията на царевичните семена могат да бъдат идентифицирани използвайки графиката за Discriminating Power. По-високите стойности на Discriminating Power указват голямо влияние на дължините на вълните в класификацията на пробите. Показателът Modeling Power показва променливи, които имат значение за описание на спектрална информация представена в известен клас от проби. Това са типични области от 0 до 1.

5) Петата стъпка включва SIMCA класификацията.

При SIMCA метода непознатите проби се класифицират към групата, за която техните параметри се вписват най-добре, т.е. разстоянието от дадената проба до центъра на модела за съответния клас е най-малко. Те могат да бъдат класифицирани в повече от една категория и може да се пресметне вероятността за принадлежност към дадена категория. Възможно е и неизвестната проба да не може да се класифицира към никоя категория.

3. Експериментални резултати

За качествен анализ на получените спектрални характеристики и за класификация на извадките със семена е използван специализирания софтуер SIMCA на програмата Pirouette Version 2.0 (Infometrics, Inc., Woodinville, WA, USA).

За получаване на по-точни резултати при класификацията са отстранени дължините на вълните в началото и края на спектралния диапазон, в които има наличие на шумове от спектрофотометъра. Това са диапазоните от 200 до 360 nm и от 950 до 1000 nm от електромагнитния спектър.

Стойностите за параметъра Interclass Distance между клас „здрави" и клас „заразени" за SIMCA модела при трите вида обработки е представен в табл.1.

Стойности за параметъра Interclass Distance Таблица 1

	изглаждане	**1-ва производна**	**2-ра производна**
Кнежа 308	5.22617	3.78896	1.15955
Кнежа 436	5.26216	4.14605	1.39389
Кнежа 613	6.06347	6.40898	2.27083
Кнежа 620	4.74254	4.31716	1.20092
26А	1.78214	1.57175	0.99729
ХМ87/136	3.58348	3.14078	1.34337
Русе 424	3.95615	3.87881	1.44120

Получените резултати от табл.1 показват, че сортове Кнежа 308, Кнежа 436, Кнежа 613, Кнежа 620, ХМ87/136 и Русе 424 са добре разделими при обработка на данните с изглаждане и първа производна, тъй като параметъра Interclass Distance има стойност над **3** [2]. За същите сортове, но при трансформация с втора производна резултатите са по-ниски. Резултатите за сорт 26А, и при трите вида преобразуване, са доста по-ниски в сравнение с останалите сортове.

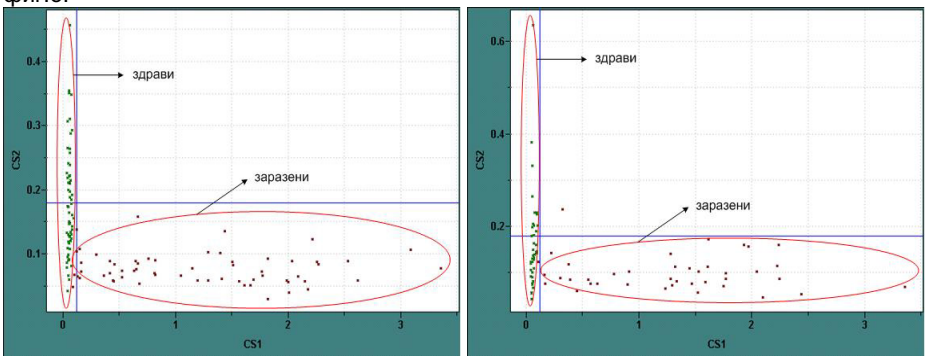
Стойностите за първите три фактора, които участват при създаването на модела за съответния клас (С1 и С2) при трите вида трансформации са представени в табл.2.

Стойности на факторите за разделимост на двата класа Таблица 2

		изглаждане		първа производна			втора производна		
		Factor1, %	Factor2, %	Factor1, %	Factor2, %	Factor3, %	Factor1, %	Factor2, %	Factor3, %
Кнежа 308	**С1**	99.72	0.20	91.61	6.72	0.96	95.08	3.39	0.49
	С2	98.89	0.74	89.22	5.92	1.99	97.15	1.35	0.49
Кнежа 436	**С1**	99.52	0.40	85.63	12.06	0.97	91.37	5.75	1.08
	С2	99.26	0.47	91.40	3.41	2.28	96.18	1.45	0.72
Кнежа 613	**С1**	99.48	0.49	88.17	10.67	0.49	93.39	4.97	0.53
	С2	99.26	0.12	92.54	3.58	1.34	97.20	1.23	0.38
Кнежа 620	**С1**	99.54	0.39	92.17	5.19	2.13	93.33	3.87	1.30
	С2	98.88	0.97	92.05	3.34	2.18	93.87	2.69	0.85
26А	**С1**	99.40	0.51	85.73	11.51	1.55	90.74	6.40	1.43
	С2	99.47	0.37	89.73	4.96	2.92	95.62	1.56	1.31
ХМ87/136	**С1**	99.82	0.12	92.60	5.12	1.37	90.34	4.91	1.90
	С2	99.28	0.57	90.68	5.35	1.69	97.15	1.66	0.46
Русе 424	**С1**	99.60	0.36	86.43	11.97	0.67	89.78	7.27	1.04
	С2	98.97	0.77	87.98	6.60	1.97	95.33	2.39	0.57

Резултатите от табл.2 показват, че най-голямо значение за разделянето на клас здрави (С1) и клас заразени (С2) има първия фактор, тъй като неговата стойност е най-близка да 100%. При класификацията с обучаващия модел и обработка на данните с изглаждане той достига максимална стойност 99,8%. При другите две обработки на данните (с първа и втора производни) се вижда, че за разделянето на класовете от значение са първите три фактора.

Параметърът Class Distances дава разстоянията от всяка проба до центровете на двата класа. SIMCA класификацията на здрави и заразени царевични семена за сорт Кнежа 308, изразена чрез този параметър и получена от модела за съответния клас при изглаждане на спектралните характеристики е показана на фиг.3.



а) б)

*Фиг. 3 SIMCA класификация на царевични семена при изглаждане на спектралните характеристики за сорт **Кнежа 308***

Двете прагови линии от фиг.3 са „критичните" стойностите за всяка обучаваща група. Всяка проба може да бъде визуално класифицирана чрез наблюдаване на позицията ѝ в подучастъците. Началните линии разделят равнината в четири квадранта. Пробите във втори квадрант принадлежат само на единия клас – в случая на клас „здрави". Пробите, попадащи в четвърти квадрант принадлежат само на другия клас – в случая на клас „заразени". Пробите в трети квадрант могат да принадлежат и към двата класа, а пробите в първи квадрант не принадлежат към нито един от двата класа.

SIMCA модела показва много добра класификация на пробите като здрави и заразени. Резултатите с пробите от тестовата група, за разпознати и неразпознати царевични семена (получени в проценти), при трите вида трансформация са показани в табл.3.

Резултати от класификацията на царевични семена със SIMCA модела Таблица 3

		изглаждане		първа производна		втора производна	
		% разпознати	% неразпознати	% разпознати	% неразпознати	% разпознати	% неразпознати
Кнежа 308	**здрави**	92.5	7.5	95	5	95	5
	заразени	92.5	7.5	92.5	7.5	97.5	2.5
Кнежа 436	**здрави**	40	60	40	60	40	60
	заразени	100	0	100	0	97.5	2.5
Кнежа 613	**здрави**	95	5	100	0	100	0
	заразени	97.5	2.5	95	5	97.5	2.5
Кнежа 620	**здрави**	97.5	2.5	97.5	2.5	87.5	12.5
	заразени	97.5	2.5	97.5	2.5	90	10
26А	**здрави**	100	0	100	0	100	0
	заразени	90	10	90	10	95	5
ХМ87/136	**здрави**	100	0	100	0	100	0
	заразени	92.5	7.5	95	5	97.5	2.5
Русе 424	**здрави**	97.5	2.5	100	0	100	0
	заразени	100	0	100	0	100	0

На базата на получените стойности от табл.3 е пресметнат общия процент на правилно и неправилно разпознати царевични семена за трите вида трансформации. Резултатите показват, че най-висок процент на правилно разпознати здрави семена (90.36%) се получава при модела с първа производна, а на правилно разпознатите заразени семена (96.43%) – при модела с втората производна.

ЗАКЛЮЧЕНИЕ

От 10-те зададени фактора, които имат отношение при съставянето на модела за съответния клас, е установено, че при обработка на данните с изглаждане е значим само първия фактор, докато при обработка с първа и втора производни – от значение са първите три фактора.

Стойностите на параметърът Interclass Distance също показват добра разделимост на класовете (здрави и заразени) за шест от изследваните седем сорта царевица. Тези резултати са на базата на изглаждане на спектрите и трансформиране с първа производна.

Обработката със SIMCA метод дава сравнително висока точност на разпознаване на заразени с Фузариоза царевични семена. Най-добри резултати са получени при модела с втора производна на спектралните данни. Средната точност за всички сортове царевични семена достига 96.43% правилно разпознати.

Целесъобразно е да се направи по-детайлен анализ на Discriminating Power и Modeling Power, за да се търсят характерните дължини на вълните за разпознаването.

ЛИТЕРАТУРА

[1] Achim Kohler, A. Skaga, G. Hjelme, H. J. Skarpeid. Sorting salted cod fillets by computer vision: a pilot study., Computers and Electronics in Agriculture, Vol.36, Issue 1, 2002, p.3-16.
[2] Atanasova, S., R. Tsenkova, R. Vasu, M. Koleva, M. Dimitrov. Identification of mastitis pathogens in raw milk by Near infrared spectroscopy and SIMCA classification method., Food Science, Engineering and Technologies, Plovdiv, 2009, p.567-572.
[3] Daskalov, P., V. Mancheva, Ts. Draganova, R. Tsonev. An approach for *Fusarium* diseased corn kernels recognition using linear discrete models., Agricultural Science and Technology, Vol.2, Number 2, 2010, p.90-95.
[4] Draganova, Ts., V. Mancheva, P.Daskalov, R. Tsonev. Wavelet based approach for Fusarium corn kernels recognition using spectral data processing., 10th IFAC Workshop on Programmable Devices and Embeded Systems (PDeS), Poland, 2010, p.19-23.
[5] Facchin S., J. Trierwieler, V. Conz, Soft sensor design: A new approach for variable selection., 2nd Mercosur Congress on Chemical Engineering.
[6] Hesti Meilina, Shinichiro Kuroki, B.M. Jinendra, Kentarou Ikuta, Roumiana Tsenkova. Double threshold method for mastitis diagnosis based on NIR spectra of raw milk and chemometrics., Biosystems Engineering, Vol.104, Issue 2, 2009, p.243-249.
[7] Pirouertte Software Manual.
[8] Атанасова, С., Хр. Даскалов, Т. Стоянчев. Приложение на спектралния анализ в близката инфрачервена област за анализ на колбаси, контаминирани с Listeria monocytogenes., Хранителна наука, техника и технологии, Пловдив, 2009, с. 561-566.
[9] БДС 607 – 73. Царевица на зърно изкупваема и за реализация., 1973.
[10] Драганова, Ц., Изследване диагностицирането на заболяването фузариоза (*Fusarium spp.*) по царевични семена чрез използване на цифрови изображения и спектрални характеристики., Докторска дисертация, 2006.
[11] Драганова Ц., Пл.Даскалов, Р.Цонев, "Алгоритъм за разпознаване заболяването фузариоза на царевични зърна чрез анализ на спектралните им характеристики.", „Научни трудове на Русенски Университет „Ангел Кънчев", Том 40, Серия 1.1, 2003, с.33-38.

За контакти:
Виолета Манчева, Катедра "Автоматика, Информационна и Управляваща Техника", Русенски университет "Ангел Кънчев", тел.:082 888684, e-mail: vmancheva@uni-ruse.bg.

Изследванията са подкрепени по договор № **BG051PO001-3.3.04/28**, „Подкрепа за развитие на научните кадри в областта на инженерните научни изследвания и иновациите". **Проектът се осъществява с финансовата подкрепа на Оперативна програма „Развитие на човешките ресурси" 2007-2013, съфинансирана от Европейския социален фонд на Европейския съюз".**

Докладът е рецензиран.