SAMOORGANIZACYJNE MODELOWANIE ROZMYTE Z WYKORZYSTANIEM METOD KLASTERYZACJI DANYCH

Tomasz Kapłon
Katedra Energetyki i Automatyzacji Procesów Rolniczych, Uniwersytet Rolniczy w Krakowie
Agnieszka Prusak
Instytut Inżynierii Rolniczej i Informatyki, Uniwersytet Rolniczy w Krakowie

Streszczenie. W pracy przedstawiono proces samoorganizacyjnego modelowania rozmytego z wykorzystaniem metod klasteryzacyjnych. Wykonano szereg modeli rozmytych typu Mamdaniego sygnału pulsacji w doju maszynowym krów i porównano ich jakość przeprowadzając analizę błędów. W procesie konstrukcji modeli stosowano algorytmy analizy skupień takie jak: K-means, fuzzy C-means, samouczące się sztuczne sieci neuronowe uczące się w trybie zwycięzca bierze wszystko i zwycięzca bierze większość. Przedstawiono ponadto własną koncepcję konfiguracji trapezowych i trójkątnych funkcji przynależności opartą na odchyleniu standardowym odległości elementów skupiska od jego centrum.

Słowa kluczowe: modelowanie rozmyte, analiza skupień, pulsacja

Wstęp

Najczęściej stosowanymi typami wnioskowania rozmytego są metoda Mamdaniego oraz Takagi-Sugeno. Metoda Mamdaniego jest szerzej wykorzystywana i akceptowana. Powodem tego jest jej większa intuicyjność oraz lepsze dopasowanie do wejść opisywanych przez człowieka. Modele rozmyte typu Mamdaniego są najczęściej budowane na podstawie wiedzy eksperta systemu. Niestety ograniczenia ludzkiego postrzegania powodują, iż możliwe okazuje się budowanie w ten sposób jedynie niskowymiarowych modeli. Przy większej komplikacji systemu, dużej ilości wejść i wyjść, eksperci nie są w stanie uzyskać dostatecznie dobrze dopasowanych modeli rozmytych. Wnioskowanie Takagi-Sugeno jest z kolei efektywniejsze obliczeniowo, lepiej dopasowane do analiz matematycznych oraz wydajniejsze dla technik optymalizacji i adaptacji. Najistotniejszą wadą tego typu wnioskowania jest to, że mogą być konstruowane jedynie za pomocą analizy danych z systemu. Modele takie stają się przez to mało zrozumiałe. Jednym z rozwiązań łączącym zalety i niwelującym wady obu typów wnioskowania jest prezentowana metoda samoorganizacji modeli rozmytych typu Mamdaniego z wykorzystaniem klasteryzacji danych. Przez model samonastrajający się trzeba rozumieć model o stałej ilości zbiorów rozmytych i stałej bazie reguł [Driankov i in.1996]. Dostrajaniu podlegają jedynie parametry funkcji przynależności i współczynniki skalowania wejść i wyjść modelu. Samoorganizujący się model rozmyty sam określa najlepszą ze względu na kryterium optymalizacyjne ilość

i postać funkcji przynależności, reguł, zbiorów rozmytych wejścia i wyjścia systemu. Modele takie są zwykle dokładniejsze w odwzorowaniu oryginalnego obiektu czy danych, ale są także trudniejsze w budowie [Piegat 1999].

Istnieją również inne metody samoorganizacji i samostrojenia modeli rozmytych np.: przekształcenie ich w rozmytą sieć neuronową, strojenie funkcji przynależności za pomocą algorytmów genetycznych, czy też metoda punktów maksymalnego błędu bezwzględnego.

Cel

Celem opracowania było dokonanie syntezy algorytmów analizy skupień oraz samoorganizujących sztucznych sieci neuronowych dla danych wielowymiarowych w celu znalezienia parametrów rozmytej funkcji przynależności typu trapezowego oraz trójkątnego. Opracowano i zaimplementowano w środowisku MATLAB program modelujący oraz przetestowano jego działanie na przykładzie odwzorowania wygenerowanego sygnału pulsacji w doju krów. Przeprowadzono także analizę błędów: średniego i jego odchylenia standardowego oraz RMSE (ang. Root Mean Square Error) - pierwiastka błędu średniokwadratowego modelowania sygnału pulsacji.

W pracy przetestowano algorytm generacji parametrów funkcji przynależności modelu pulsacji w aparacie udojowym dla krów przy zmiennych parametrach takich jak:

– rodzaj funkcji przynależności: trójkątna lub trapezowa,
– użyty algorytm analizy skupień: K-means, C-means, samoorganizujące się sztuczne sieci neuronowe (SOM SNN) w trybie uczenia Winner Takes All (WTA) oraz Winner Takes Most (WTM).

Metodyka

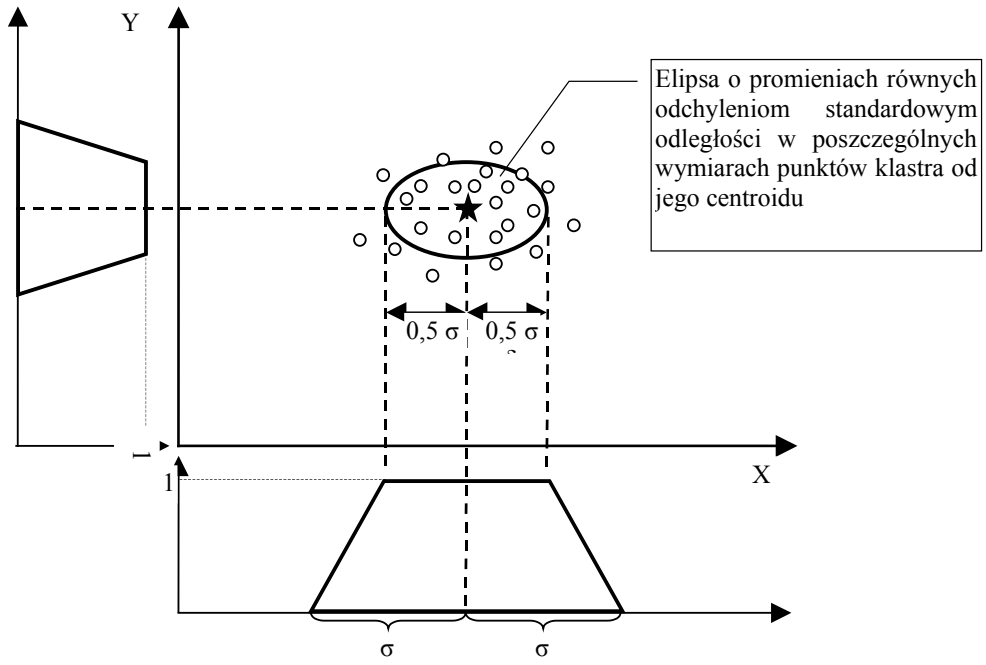
W badaniach konstruowano samoorganizacyjne rozmyte modele typu Mamdaniego metodami klasteryzacyjnymi. Zaproponowano własną koncepcję konfiguracji użytych funkcji przynależności (rys. 1). Kluczowym parametrem konfiguracyjnym jest rozmiar odnalezionego skupiska, opisywany parametrem σ, który dla każdego z nich, w każdym wymiarze jest inny i równy co do wartości odchyleniu standardowemu odległości między centroidem a obiektami do tego klastra zakwalifikowanymi.

W opracowywanych modelach systemów stosowano najprostsze bazy reguł. Dla systemu typu MISO o dwóch wejściach x_1, x_2 i jednym wyjściu y baza reguł wyglądałaby następująco:

Reguła 1: **IF** $(x_1=A_1)$ **AND** $(x_2=B_1)$ **THEN** $(y=C_1)$
Reguła 2: **IF** $(x_1=A_2)$ **AND** $(x_2=B_2)$ **THEN** $(y=C_2)$
⋮
Reguła N: **IF** $(x_1=A_N)$ **AND** $(x_2=B_N)$ **THEN** $(y=C_N)$,

gdzie:

N – liczba funkcji przynależności. W konstruowanych modelach wartość ta wynika z ilości wykrytych w analizie skupień klastrów.



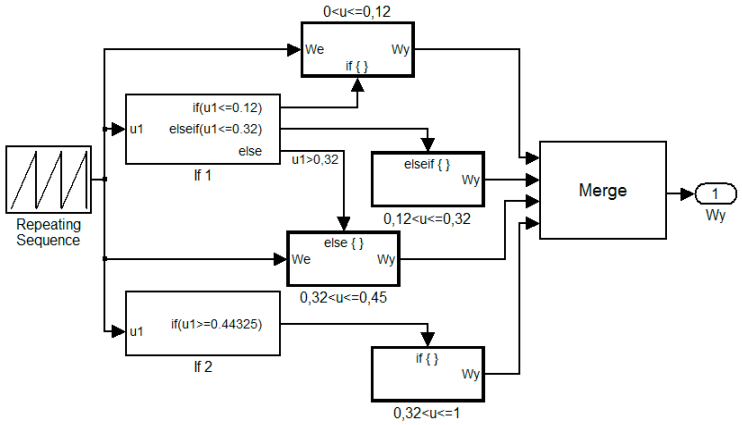
Źródło: opracowanie własne

Rys. 1. Generacja trapezowych funkcji przynależności z parametrów jednego klastra dla danych dwuwymiarowych

Fig. 1. Generation of trapeze-membership functions from parameters of single cluster for two dimension data

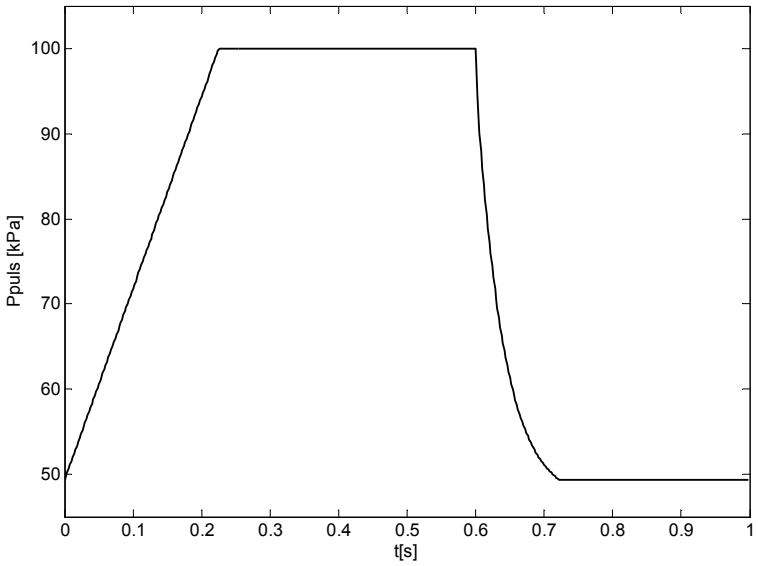
Sygnał pulsacji, stanowiący dane uczące i testujące, tworzące macierz [t, P_{puls}], wygenerowano na podstawie symulacji komputerowej literaturowego modelu zmian ciśnienia bezwzględnego w komorze międzyściennej kubka udojowego w aparacie udojowym dla krów (rys. 2). Typy funkcji oraz odpowiednie współczynniki dobrano na podstawie wyników badań dotyczących charakterystyk statycznych gum strzykowych, modelowania matematycznego ich ruchu, zmian ciśnienia, objętości komory międzyściennej i podstrzykowej kubka udojowego oraz modelowania podciśnienia w kolektorze aparatu udojowego dla krów [Kupczyk 1988, Juszka i in. 2005, Juszka, Kapłon 2010]. Zestaw uczący i testujący zawierał po 500 par danych odnośnie czasu t [s] i odpowiadającego mu ciśnienia bezwzględnego P_{puls} [kPa] w komorze międzyściennej kubka udojowego w ciągu jednej sekundy. Dwuwymiarowe dane pozwoliły na budowę modelu typu SISO (ang. Single Input Single Output), którego wejściem był czas, wyjściem zaś ciśnienie.

Kierując się wymaganiami algorytmów analizy skupień oraz SSN, oba zestawy danych zostały znormalizowane do przedziału [0, 1] (rys. 3). W wyniku przeprowadzonych badań empirycznych i wykonaniu analizy błędów określono następujące zakresy analizy: od 5 do 95, w odstępach co 10 klastrów dla klasteryzacji metodami K-means i C-means oraz od 3 do 12 w odstępach co 1 dla samouczących się SSN przy metodach uczenia WTA i WTM.



Źródło: badania własne

Rys. 2. Model sygnału pulsacji w aparacie udojowym dla krów zrealizowany w programie MATLAB Simulink
Fig. 2. Signal of pulsation in milking cluster in MATLAB Simulink generator

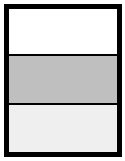


Źródło: badania własne

Rys. 3. Dane uczące odwzorowujące sygnał pulsacji w aparacie udojowym dla krów
Fig. 3. Training data of pulsation signal in milking cluster

Wybrano inny zakres dla algorytmów K-means i C-means niż dla SSN, gdyż duża ilość neuronów powodowała efekt uczenia się przez nie poszczególnych danych, co powodowało przybywanie neuronów martwych - nieuczących się w żadnym kroku. Duża ilość neuronów skutkowała również wydłużeniem czasu uczenia się sieci.

Wyniki modelowania przedstawiające:



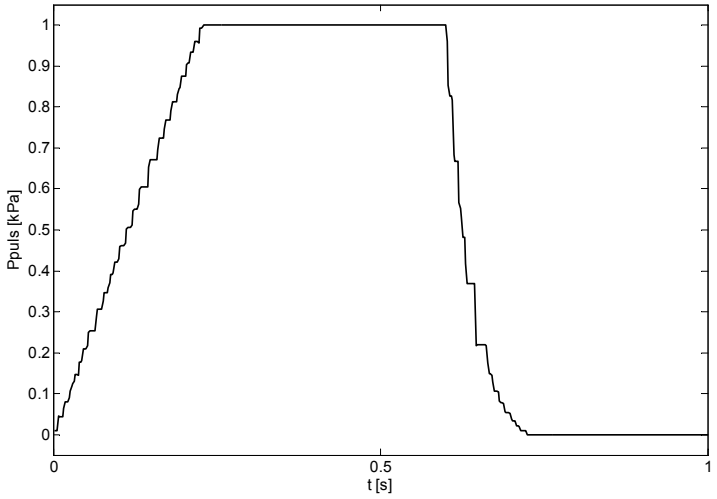
- błąd RMSE dla danych testujących,
- błąd średni [%],
- odchylenie standardowe błędu średniego [%] zamieszczono w tabeli 1.

Tabela 1. Błędy modelowania sygnału pulsacji
Table 1. Errors of pulsation signal modeling

Algorytm klasteryzujący	Funkcja przynależności	Ilość klastrów									
		5	15	25	35	45	55	65	75	85	95
K-means	Trap.	0,064	0,027	0,019	0,015	0,014	0,010	0,017	0,020	0,028	0,023
		4,015	1,498	0,888	0,667	0,633	0,433	0,733	0,753	0,915	0,999
		4,947	2,252	1,692	1,353	1,288	0,901	1,536	1,871	2,604	2,067
	Trój.	0,076	0,045	0,030	0,018	0,029	0,011	0,011	0,025	0,008	0,075
		4,223	2,026	1,318	0,819	1,048	0,469	0,459	1,049	0,344	1,725
		6,359	4,023	2,744	1,602	2,660	0,967	1,021	2,309	0,740	7,279
C-means	Trap.	0,152	0,032	0,016	0,014	0,095	0,009	0,009	0,009	0,008	0,007
		10,305	1,690	0,825	0,622	2,309	0,414	0,361	0,372	0,312	0,288
		11,172	2,754	1,427	1,236	9,274	0,818	0,803	0,857	0,735	0,632
	Trój.	0,149	0,031	0,018	0,013	0,011	0,011	0,010	0,008	0,008	0,007
		10,142	1,624	0,948	0,622	0,506	0,423	0,388	0,326	0,311	0,283
		10,945	2,585	1,547	1,185	1,023	1,002	0,953	0,702	0,741	0,659
		Ilość klastrów									
		3	4	5	6	7	8	9	10	11	12
SOM WTA	Trap.	0,088	0,160	0,075	0,065	0,091	0,074	0,301	0,339	0,296	0,294
		6,220	11,528	4,217	3,841	4,715	3,733	20,483	25,750	20,078	19,569
		6,169	11,119	6,236	5,209	7,748	6,386	22,022	22,032	21,705	21,962
	Trój.	0,087	0,166	0,075	0,064	0,101	0,058	0,110	0,104	0,048	0,049
		6,179	12,098	4,154	3,881	4,903	3,454	5,212	4,846	2,852	2,752
		6,060	11,402	6,218	5,102	8,826	4,658	9,676	9,246	3,895	4,022
SOM WTM	Trap.	0,369	0,376	0,296	0,256	0,243	0,233	0,174	0,161	0,301	0,302
		34,22	32,10	21,53	14,43	16,23	15,50	10,08	9,18	21,52	21,42
		13,78	19,58	20,38	21,17	18,13	17,38	14,23	13,18	21,10	21,26
	Trój.	0,363	0,365	0,291	0,255	0,238	0,243	0,158	0,130	0,183	0,121
		33,51	31,44	21,01	14,58	15,88	16,25	9,14	7,18	10,27	6,27
		13,943	18,502	20,160	20,937	17,808	18,110	12,839	10,849	15,126	10,406

Źródło: badania własne

Analizując tabelę można stwierdzić, że najlepiej zadany sygnał aproksymowały modele utworzone z użyciem klasteryzacji K-means i C-means. Najsłabiej radził sobie algorytm z SOM przy uczeniu WTM. Najmniejszy błąd uzyskano dla modelu z 95 trójkątnymi funkcjami przynależności, których parametry uzyskane zostały przy pomocy algorytmu C-means. Błąd RMSE wynosił 0,007, zaś błąd średni 0,283% przy odchyleniu standardowym 0,659%. Odwzorowanie sygnału pulsacji przez ten model przedstawia rys. 4.



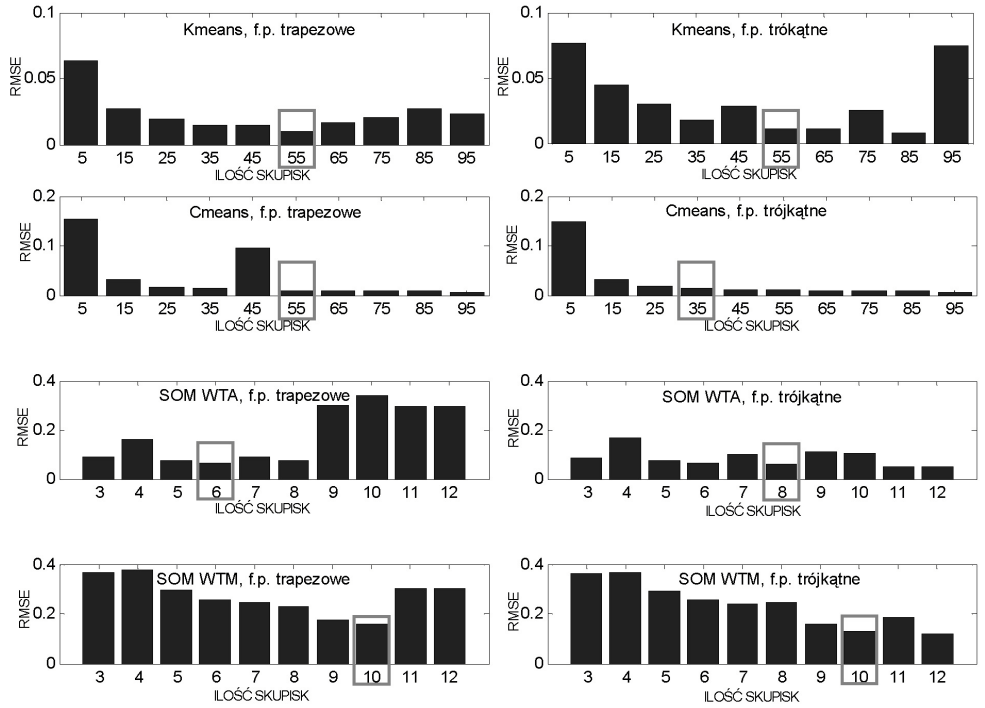
Źródło: badania własne

Rys. 4. Odpowiedź najlepszego modelu na testujący wektor wejściowy
Fig. 4. An answer of the best fuzzy model output for a testing input vector

W celu porównania proponowanej metody ze znanymi i sprawdzonymi metodami modelowania rozmytego wykonano również za pomocą algorytmu MATLABa ANFIS [Mrozek 2006] modele typu Takagi-Sugeno dostrojone poprzez przekształcenie w sieć neuro-rozmytą. Uzyskane najlepsze wyniki obarczone były błędem na poziomie:

1. RMSE = 0,005 - dla 60 trapezowych funkcji przynależności,
2. RMSE = 0,006 - dla 100 trójkątnych funkcji przynależności.

Istotnym parametrem modeli rozmytych jest ilość funkcji przynależności opisujących zmienną rozmytą. Duża ich liczba powoduje zbytnie skomplikowanie modelu, szczególnie przy wielu jego wejściach i wyjściach. Zwiększa się również jego złożoność obliczeniowa, a techniki dostrajania go okazują się niewydolne. Modele takie mogą również odwzorować system zbyt dokładnie, co oznacza pozbawienie modelu zdolności „generalizacji" danych. Dlatego właśnie konieczny jest dodatkowy krok algorytmu, polegający na kompromisowym wyborze najlepszego modelu ze względu na jego szybkość i dokładność. Należy założyć pewien próg poprawy błędu modelownia w kolejnym kroku zwiększania ilości funkcji przynależności. Jeżeli nie zostaje on przekroczony należy za najlepszy wybrać model poprzedni - prostszy. Graficznie przedstawia to rysunek 5.



Źródło: badania własne

f.p. – funkcje przynależności

Rys. 5. Zmiana wartości błędu modelowania RMSE wraz ze wzrostem liczby skupisk z zaznaczonymi najlepszymi modelami

Fig. 5. Change of the RMSE modelling error value with a cluster number increase and the best models marked

Podsumowanie

Przedstawiona w pracy metoda samoorganizacji modeli rozmytych z wykorzystaniem algorytmów analizy skupień pozwoliła na sformułowanie poprawnych modeli sygnału pulsacji w doju krów. Proponowana, oryginalna metoda konfiguracji trójkątnych i trapezowych funkcji przynależności zapewniła dla modelowanego sygnału wyniki porównywalne z dużo bardziej skomplikowanymi i złożonymi obliczeniowo metodami strojenia, opartymi na modelach neuro-rozmytych. Dzięki zastosowaniu najprostszych funkcji przynależności uzyskane modele są szybkie obliczeniowo i łatwe w implementacji w sterownikach i regulatorach wykorzystujących logikę zbiorów rozmytych.

Bibliografia

Driankov D., Hellendoorn H., Reinfrank M. 1996. Wprowadzenie do sterowania rozmytego. Wydawnictwo Naukowo-Techniczne, Warszawa. ISBN 83-204-2030-X.

Juszka H., Lis S., Tomasik M. 2005. Modelowanie i sterowanie rozmyte aparatem udojowym. Problemy Inżynierii Rolniczej. Nr 4(50). Warszawa. s. 57-64.

Juszka H., Kapłon T. 2010. Modelowanie i symulacja komputerowa częstotliwości pulsacji w aparacie udojowym. Inżynieria Rolnicza. Nr 4(122). s. 99-106.

Kupczyk A. 1988. Model dynamiki zmian podciśnienia w aparacie udojowym dojarki z rozdzielnym transportem mleka i powietrza. Cz. I. Model dynamiki zmian podciśnienia w kubku udojowym. Roczniki Nauk Rolniczych. t. 78 - C - 2. Warszawa. s. 173-181.

Mrozek B. 2006. Projektowanie regulatorów rozmytych w środowisku MATLAB-Simulink, Pomiary Automatyka Robotyka. Nr 11. s. 5-12.

Piegat A. 1999. Modelowanie i sterowanie rozmyte. Akademicka Oficyna Wydawnicza EXIT. Warszawa. ISBN 83-87674-14-1.

SELF-ORGANIZING FUZZY MODELLING WITH DATA CLUSTERING METHODS

Abstract. The study presents the process of a self-organizing fuzzy modelling by data clustering method. The Mamdani type models of pulsation signal in a cow machine milking cluster were made and error analysis was carried out. Different clustering algorithms like: K-means, fuzzy C-means, Self-Organizing Maps in Winner Takes All or Winner Takes Most training methods were used in the process of the model construction. Moreover, an original idea of trapeze and triangle-membership functions in the fuzzy variables based on a standard deviation of distance data in a cluster from its centroid were presented.

Key words: fuzzy modelling, cluster analysis, pulsation

Adres do korespondencji:
Tomasz Kapłon; email: tomaszkaplon@op.pl
Katedra Energetyki i Automatyzacji Procesów Rolniczych
Uniwersytet Rolniczy w Krakowie
ul. Balicka 116B
30-149 Kraków